Essays in Global Games and Political Economy

Essays in Global Games and Political Economy

A dissertation presented

by

Thomas Russell Gole

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

September 2013

Dissertation Advisor:
Professor Alberto Alesina

Author:
Thomas Russell Gole

Essays in Global Games and Political Economy

Abstract

This dissertation consists of three essays concerned with coordination, cooperation and the governance of institutions.

The first chapter analyzes the effect of coordination incentives on committee decision-making. When members of a selection committee have incentives to agree with each other, they over-weight public information; this generates statistical discrimination. We test this hypothesis using a novel field experiment — a large debate tournament in which judges are randomly assigned to committees that decide results — and find that judges with greater desire to coordinate are more likely to vote for teams with better past records. To understand the magnitude and implications of these estimated effects, we then develop and estimate a structural model in which committee members with incentives to cooperate receive noisy signals of candidate quality. Our results confirm that public information can cause committees to coordinate on weaker candidates.

The second chapter considers the governance challenges posed by the developing technology of geoengineering. We argue that geoengineering may constitute a "free-driver" problem, in which the country or actor that suffers most from climate change free-drives the global level of geoengineering. The chapter presents a simple model of free-driving and identifies the parameters that govern whether a problem is one of free-riding or free-driving. We apply this model to geoengineering, by providing a qualitative synthesis of the literature on climate change damages, then by using estimates of regional climate damage

heterogeneity from the RICE model (Nordhaus, 2010). The result is a first-pass attempt at quantitatively identifying which regions are most likely to be in favor of geoengineering, and which against. It appears that free-driving is a serious possibility, but there is significant space for effective negotiation.

The third chapter combines a general election model in which candidates have policy preferences with a primary election process which takes the form of a citizen-candidate model. We use this to establish conditions under which both models are well-behaved, and then characterize equilibria. Divergence of proposed platforms within the party primary is common, and in equilibrium candidates more extreme than their party median may not only stand for, but win both the primary and general elections.

Contents

List of Tables

List of Figures

Acknowledgments

Over the past four years that have gone into this dissertation, I have been blessed with a remarkable set of opportunities to work with fantastic people and learn a great deal about economics. This could not have been possible without the generosity of friends, colleagues and advisors.

Firstly, I am grateful for the guidance of my thesis committee: Alberto Alesina, who has consistently been willing to help and to converse freely on a wide range of topics; Andrei Shleifer, who has always been willing to give me useful and frank advice over the past four years; and David Laibson, who has been generous with his time. Their advice throughout my graduate school experience has been invaluable. I am also intensely grateful to Brenda Piquet for all of her efforts in my aid.

I have been lucky to have had two great co-authors, Simon Quinn and Jisung Park. It has been a pleasure to work with each of them, and I hope to continue to do so in the future. I would also like to thank everyone who has given my co-authors and me useful feedback: Philippe Aghion, Debopam Bhattacharya, Steve Bond, Ian Crawford, Markus Eberhardt, Chris Erskine, Oliver Hart, Cecilia Heyes, Richard Holden, Christopher Hope, David Keith, Clare Leaver, William Nordhaus, Erin O'Brien, Eric Maskin, Andy Parker, Aureo de Paula, Ken Shepsle, Matt Shum, Russell Smyth, Kathy Spier, Marty Weitzman and Ken Wolpin. Thanks must also go to seminar audiences at the Australian National University, the Queensland University of Technology, the University of New South Wales, the University of Oxford and the University of Queensland, as well as the Harvard International, Environmental, Political Economy, Industrial Organization and Organizational Behavior lunches.

During the past four years I have also had the pleasure of working with and learning from a series of amazing people: Julie Mortimer and Chris Conlon; Raj Chetty and Gita

Gopinath; Greg Mankiw, David Johnson and the Ec10 team; Alberto Alesina, Marty Weitzman and Jeff Frieden whose courses I had the pleasure of teaching; Michael Porter and the Institute for Strategy and Competitiveness; Amartya Sen; everyone I worked with at the IMF on the October 2012 Global Financial Stability Report, in particular Laura Kodres, Tao Sun and the Global Stability team; Anton Dobronogov and the Africa PREM 3 team at the World Bank; and finally John Briscoe and the Harvard Water Security Initiative, in particular the Murray-Darling team for the 2012 Water Federalism Project.

I would also like to acknowledge the support of the Frank Knox Memorial Fellowship. The Knox Fellowship is more than just financial support: it is also a remarkable group of young people, many of whom have become the closest of friends. I also greatly appreciated the support of the Terence M. Considine Fellowship in Law and Social Sciences. I must also thank the support and friendship of everyone at the Committee on General Scholarships: Dean Margot Gill, Dina Moakley, Rebecca Lock, Lisa Bruzzese and Gabriela Bacares.

To friends and colleagues over the past four years, be it from the football club, the economics department or the wider Harvard community, you know who you are, and I appreciate your friendship and your support. To Aurélie Ouss, Alex Peysakhovich, Jesse Schreger and Oren Ziv in particular, your friendship has meant a very great deal to me over the past four years.

As with all things in my life, I owe a very great debt of gratitude to my family: my siblings Tim and Hobia, and my parents. Everything I am, have done and will ever do is thanks to them.

Finally, the greatest thanks of all must go to Kavita, who moved across the world for me.

Chapter 1: Committees, Statistical Discrimination and Global Games: Structural Evidence from a Randomized Field Experiment[1]

1.1 Selection committees as beauty contests

Committees matter. In private corporations and in government agencies, most important decisions are taken through committee voting. The rules that govern such voting can be critical for decisions as diverse as the hiring of job candidates (Goldin and Rouse, 2000), the setting of monetary policy (Riboni and Ruge-Murcia, 2011; Jung, 2011; Havrilesky and Schweitzer, 1990; Gildea, 1990), and the determinations of courts of law (Iaryczower *et al.*, 2013; Iaryczower and Shum, 2012; Blanes i Vidal and Leaver, 2013; Levy, 2005).[2] But the committee is also a very difficult institution to study. It is a difficult institution to model theoretically, because of the complexity of different incentives at play: a committee will typically have a collective objective, but committee members also hold individual preferences.

[1]Co-authored with Simon Quinn (Department of Economics and the Centre for the Study of African Economies, University of Oxford).

[2]The Supreme Court of the United States has recognized this fact for many years. Supreme Court Justice Tom C. Clark said this in 1959: "Ever since John Marshall's day [*i.e.* the early 1800s] the formal vote begins with the junior Justice and moves up through the ranks of seniority, the Chief Justice voting last. Hence the juniors are not influenced by the vote of their elders!" (Clark, 1959, page 50).

It is also a difficult institution to evaluate empirically: data on committee decisions rarely includes detailed information on individual members' perceptions, and exogenous variation in participants' incentives is very uncommon.

In this paper, we measure the effect on committee outcomes of individual members' preferences for agreement. We use a randomized field experiment with a novel design, in which participants are repeatedly assigned to different three-member committees to assess competing teams in a tournament. Our hypothesis is simple: *when members of a selection committee have incentives to agree with each other, they over-weight public information in reaching their decisions*. In effect, a selection committee can operate like a Keynesian beauty contest, where participants worry not only about their own perceptions of the candidates, but also about the perceptions of others.[3]

To test this claim, we exploit quasi-random variation in committee outcomes to generate exogenous variation in committee members' preferences for future agreement. We argue that participants who have just dissented from their peers have a stronger preference for future agreement than they otherwise would. We show a large, significant and robust effect of such past dissent on the probability of voting for a pre-tournament favorite: that is, an effect on participants' weighting of public information. We distinguish this effect from potential learning about signal quality by controlling for a finer ex post measure of disagreement that judges receive after each vote (see Section 1.2).

For this reason, we argue that the committee — an institution omnipresent in hiring

[3]In Chapter 12 of *The General Theory of Employment, Interest, and Money*, Keynes (1936) famously said this: "...professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view." Costa-Gomes and Crawford (2006) quote the same metaphor in motivating their experimental study of level-k thinking.

decisions — can itself act as a mechanism by which disadvantaged groups suffer statistical discrimination. Researchers have characterized statistical discrimination for at least 40 years, with increasing attention paid to the different forms that the phenomenon can take. Phelps (1972) famously described such discrimination in terms of employers' conditional expectations of employee productivity, where "color or sex is taken as a proxy for relevant data not sampled"; Arrow (1973) embedded the same idea in a dynamic model, arguing that discrimination can arise endogenously in response to discriminatory beliefs, generating path dependence (see also Coate and Loury (1993), Moro and Norman (2004) and Fryer (2007)). Empirical research has generally been concerned to test for the existence and magnitude of discrimination. Recent work has tested the effect of variation in the characteristics of assessors — for example, variation in gender (Bagues and Esteve-Volart, 2010; Beck *et al.*, 2012) or race (Anwar *et al.*, 2012; Price and Wolfers, 2010). Other results have exploited exogenous variation in candidate identity; this has included both natural variation in the observability of candidate characteristics (Goldin and Rouse, 2000; Lavy, 2008) and randomized variation in the attributes of 'fake' applicants (Bertrand and Mullainathan, 2004; Hanna and Linden, 2012).

But very little work — theoretical or empirical — has sought to discover the underlying incentive mechanisms for statistical discrimination; that is, the institutional features that might encourage a decision-maker to place more emphasis upon his or her *a priori* beliefs about a candidate, rather than to assess the candidate based upon his or her individual merits.[4] This presents an important area for further learning on statistical discrimination, and one that is ripe for both theoretical and empirical insight.

We make several contributions. First, we implement a *novel experimental design* in which

[4]List (2004, page 52) makes a very similar observation about the challenge of distinguishing statistical discrimination from pure prejudice: "An important lesson learned from the vast literature on discrimination is that data availability places severe constraints on efforts to understand the nature of discrimination, forcing researchers to speculate about the source of the observed discrimination."

participants are randomly assigned to different committees, and committees are randomly assigned to assess different candidates. This kind of design suggests many possibilities for new experimental research on aspects of group behavior (see also Fafchamps and Quinn (2012) and Boudreau *et al.* (2013)).

Second, we provide the first *empirical evidence* that a preference for coordination can generate statistical discrimination in committee decision-making. This complements the large empirical literature on statistical discrimination, by showing a new mechanism by which such discrimination can arise. Further, the results provide empirical support for several theoretical assertions about committee behavior. For example, Levy (2007) models the effect of transparency on committees, arguing that committee members' career concerns can encourage members to 'conform to preexisting biases'.[5]

Third, we develop a *new structural methodology* for estimating Bayesian games where players receive correlated signals. Previous work on structural estimation of Bayesian games (for example, de Paula and Tang (2012)) has focused on identifying the sign of coordination preferences non-parametrically. Such identification strategies generally rely upon the assumption that, conditional on covariates observable to the researcher, players' signals are independent. We show how exogenous variation in player preferences may be exploited to point-identify the magnitude of players' preference for coordination, even under correlated signals. We achieve this result by making several extensions to recent theoretical results on uniqueness in discrete global games. Our consequent structural estimates allow us to interpret our experimental results in terms of underlying preferences; they also allow us to predict behavior under a counter-factual in which players hold no preference for coordination.

[5]Visser and Swank (2007) consider a model in which committee members prefer to conceal disagreement, in order to preserve their reputation.

Fourth, we take a novel approach to *structural model validation*. We used our structural model to make testable predictions for a later iteration of the experiment, and registered the model and predictions at the J-PAL Hypothesis Registry.[6] Registration is increasingly important in randomized controlled trials, where which it is seen as a valuable method for committing to specifications and hypotheses before experiments are run (see, for example, Casey *et al.* (2012)). To our knowledge, hypothesis registration has not previously been used for out-of-sample validation of a structural model. However, in a context like this — where we use a structural model for the analysis of an experiment — registration is a credible and transparent method for testing out-of-sample model validity.

Three sections follow. Section 1.2 describes our experiment and presents regression results. Section 1.3 develops and estimates a new structural model. We conclude in Section 1.4.

1.2 A novel field experiment

For causal inference on the role of coordination preferences, researchers ideally need an experimental context with several quite peculiar features. First, participants should be assigned to committees randomly, so that observed behavior cannot be attributed to endogenous committee formation. Second, participants should face random or quasi-random shocks to their preferences over coordination, and these shocks should be asymmetric between different committee members; this allows researchers to measure the effect of coordination preferences, distinct from the effect of other information that a committee may receive. Third, committee voting procedures should be specific and standardized across different committees. Finally, researchers should ideally study a situation where payoffs are meaningful and where participants are very familiar with the context and the committee

[6]The J-PAL Hypothesis Registry allows for time-stamped registration of hypotheses or predictions for randomized controlled trials. The Registry, including our predictions, is online at `http://www.povertyactionlab.org/Hypothesis-Registry`.

protocols — that is, a 'natural field experiment' (Harrison and List, 2004).[7]

We study a novel experiment that has all of these features. Almost no previous research has considered committee voting in a randomized field context.[8] In this way, the present experimental context provides a new method for testing committee interactions.

1.2.1 The World Schools Debating Championships

The World Schools Debating Championships are an annual debate tournament between high school students. Debaters are drawn from around the world to represent their countries; each nation is entitled to one team in the competition.[9] The Championships are the premiere international debate tournament for school students.[10] We study the Championships held in 2010 (in Doha, Qatar), in 2011 (in Dundee, Scotland) and in 2012 (in Cape Town, South Africa). A total of 66 countries competed at these three tournaments, of which 39 countries participated at all three.[11]

Each debate pitches one national team against another; teams are randomly assigned

[7]Harrison and List (2004) describe a natural field experiment as a context "where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are in an experiment" (page 1014). In our context, participants were told that data would be collected about their decisions, and that this might be used for academic research in economics.

[8]For example, Fafchamps and Quinn (2012) report initial results from the first randomized field experiment to form committees of entrepreneurs — but their emphasis is on peer effects among participants, rather than on the effect of preferences for agreement. Similarly, Boudreau *et al.* (2013) randomly assign medical researchers to information-sharing sessions, to study effects on future collaboration. In related work, other researchers have considered experiments played on networks, in which network structure is varied to test its effects on coordination and diffusion: see, for example, Boosey (2011), Centola (2010) and Centola (2011).

[9]The Championships started in 1988, and have now run 24 times. Four nations have been represented at all of those tournaments: Australia, Canada, England and the United States.

[10]The Championships typically attract media attention, both in their host countries and elsewhere. For example, *Team Qatar*, a documentary about the Championships by American director Liz Mermin, made its world premiere at the Tribeca Film Festival in 2009.

[11]Extensive information on the Championships — including on the rules and history of the tournament — is available at the official website: `http://www.schoolsdebate.com/`.

to argue either for or against a controversial idea.[12] The Championships comprise both Preliminary Rounds and Finals Rounds. In the Preliminary Rounds, each nation competes against eight randomly-drawn opponents. These eight debates occur across four days: Rounds 1 and 2 on the first day, Rounds 3 and 4 on the second day, and so on. The top 16 teams then progress to the Finals Rounds, a series of five knock-out debates culminating in the Grand Final.[13] Our analysis focuses exclusively on data from the Preliminary Rounds.[14]

Judging committees: The winner of each debate is determined by a committee of three judges. Together, this committee is required to decide which team has argued more persuasively.[15] Judges assesses the debate separately, assigning points to speakers based on the categories of 'style', 'content' and 'strategy'.[16] Each judge is required to complete a ballot, in which he or she records speaker points and decides the winner of the debate; judges may not award a tie. An example ballot is provided in Appendix A.3, along with further explanation on the marking categories. The debate is won by whichever team wins two or three of the judges; committee outcomes can therefore be either 'unanimous' (3-0) or 'split' (2-1).

Critically, judges are not allowed to communicate with each other (or with the competitors) until *after* making their decisions.[17] Having made their decisions, the three judges

[12]For example, the 2010 Championships began with a debate on the proposition "That we should support military intervention in Somalia"; the same Championships ended with a Grand Final debate on the proposition "That governments should never bail out big companies".

[13]For example, in 2010, Canada won all eight of its Preliminary Round debates, defeating (in order) Bangladesh, Botswana, Thailand, Argentina, Namibia, South Korea, Palestine and Pakistan. In the Finals Rounds, the team then defeated Ireland, New Zealand, Singapore and England (in order), to become the World Schools Debating Champions.

[14]We limit our sample in this way because judge assignment for the Finals Rounds is not random.

[15]That is, the committee does *not* decide whether it agrees or disagrees with the proposition being debated. Judges' personal views about the issue under debate are not considered to be relevant for the assessment of which team has better argued its case.

[16]A comprehensive explanation of these categories is available at `http://www.schoolsdebate.com/`.

[17]Judges are seated apart. There is no evidence of judges trying to 'cheat' by looking at each other's notes;

then leave the room to confer; judges may not change their decisions after leaving the room. Having discussed the debate together, the committee returns to the room; one judge announces the committee's result, and gives a brief justification for the committee's decision. Teams and their coaches are then encouraged to speak separately with the judges; at this point, there is a strong emphasis on constructive feedback.

In this way, the Championships provide an ideal field experiment in which to study the consequences of committee voting for the expression of members' private information. Literature on committee voting tends to emphasize two distinct roles for committee processes: (i) aggregation of disparate information and (ii) communication/persuasion between committee members (see, for example, Austen-Smith and Feddersen (2009, 2006) and Feddersen and Pesendorfer (1996)). Communication/persuasion is an important aspect of many real-world committees. However, the effect that we study is an effect on committee members' expression of private information; that is, an incentive that may discourage committee members from sharing their private perceptions in a completely informative way. It is critical for our field experiment that judges cannot communicate before they vote, because this allows us to isolate the effect of past dissent on each judge's *individual* decision.

The judges: Judges come from around the world to participate in the Championships; across the three tournaments studied, a total of 49 nations were represented on various judging committees. Judges are volunteers, and most are required to pay their own travel and accommodation expenses to participate.[18] Most judges are young and highly educated. In 2012, the median age of the judges being studied was 27. All judges have completed secondary school; 70% have completed an undergraduate degree, and 40% have completed a postgraduate degree (primarily in social sciences and humanities — for example, in politics,

indeed, there are strong norms at the Championships against such behavior. Judges are also discouraged from allowing their facial expressions or body language to indicate their views on the debate.

[18]Some nations subsidize their judges' expenses. Additionally, in 2010, the host organization (QatarDebate) paid the travel and expenses of 37 experienced judges, in order to ensure that sufficient judges were able to participate in the tournament.

English, law, economics or history).[19]

Random assignment: In total, we study 603 committees across the 2010, 2011 and 2012 tournaments; Table 1.1 summarizes.[20]

Table 1.1: ***Number of committees by tournament***

	tournament			
round	2010	2011	2012	**total**
1	28	24	24	76
2	28	24	24	76
3	27	24	24	75
4	27	24	23	74
5	27	24	24	75
6	27	24	24	75
7	28	24	24	76
8	28	24	24	76
total	220	192	191	**603**

Judges were assigned to committees randomly (using a computer), and judges knew this.[21] This assignment was subject to several constraints, designed to improve the 'balance' of the randomization.[22] Most importantly, each committee comprised one 'class 1' judge (most experienced/competent), one 'class 2' judge and one 'class 3' judge (least experi-

[19]This information is drawn from an online survey of judges that we ran after the 2012 Championships. Of the 222 judges who participated in the three tournaments we study, 174 answered the online survey (*i.e.* about 78%). We do not use this survey data for any substantive analysis, but we feel that it provides a reasonable description of judge characteristics.

[20]In the 2010 tournament, we had 28 debates in each round. However, one of the authors (Quinn) was one of the two Chief Adjudicators of that tournament, and was required to judge on four committees (in rounds 3, 4, 5 and 6); we have dropped those committees from the dataset. In 2010, we needed four extra debates — in a notional 'Round 0' — in order to ensure that the draw was balanced between the 57 competing teams. We have dropped these committees. In the 2012 tournament, we had 24 debates in each round. However, one judge fell ill during a debate in round 4, and was required to withdraw from the committee decision; we have also dropped that committee.

[21]The computer code was written in Stata, and is available on request.

[22]For a general discussion of the issue of balance and randomization in field experiments, see Bruhn and McKenzie (2009).

enced/competent). These classes were assigned subjectively by the tournament organizers, to ensure a balance of judging experience across different committees. Second, each committee included at least one man and one woman (58% of all judges being male).[23] Third, we limited cases of judges seeing the same team more than once in the same tournament, and no judge was allowed to assess his or her own team.[24] Fourth, because the Preliminary Rounds were usually divided between different venues (often high schools), we often needed to assign pairs of judges to committees together in two debates on the same day. Figure 1.1 shows the consequent network of committees that we observe.[25] Nodes represent judges (with the three rows respectively showing classes 1, 2 and 3, and node size reflecting the number of debates assessed), and each edge shows that two judges have worked together on the same committee. The figure emphasizes that each judge had a wide variety of peers.

Pre-tournament rankings: Teams are ranked before each tournament. On the basis of this ranking, a random draw determines each team's position in the draw.[26] Pre-tournament ranking is necessary so that each team is drawn against opponents of a range of different qualities; *i.e.* so that a team does not face a disproportionate number of very strong teams in its Preliminary Rounds, nor a disproportionate number of weaker teams.[27] Teams are ranked on the basis of their performance in the Preliminary Rounds of the three previous

[23]We were required to relax this constraint four times: in 2010, we allowed one all-male committee, in 2011, we allowed two all-female committees, and in 2012, we allowed one all-male committee.

[24]Readers may nonetheless be concerned about the incentive for judges to make decisions that help the position of their national team in the overall standings. There are several reasons that we do not believe that this is a common phenomenon. First, the complexity of the tournament often makes it difficult for participants to know how particular results may or may not assist other competing teams. Further, this kind of strategy could easily backfire through substantial reputational harm both to the individual judge and to the tournament as a whole — and, as we note shortly, such reputation appears to matter for participating judges.

[25]The figure is generated using the software package *Pajek*: see De Nooy *et al.* (2011).

[26]This random draw is filmed and made available online; for example, the draw videos from the 2010 tournament are available at `http://www.youtube.com/user/WSDC2010`.

[27]Many tournaments use a similar approach for seeding a random draw — including, for example, the FIFA World Cup.

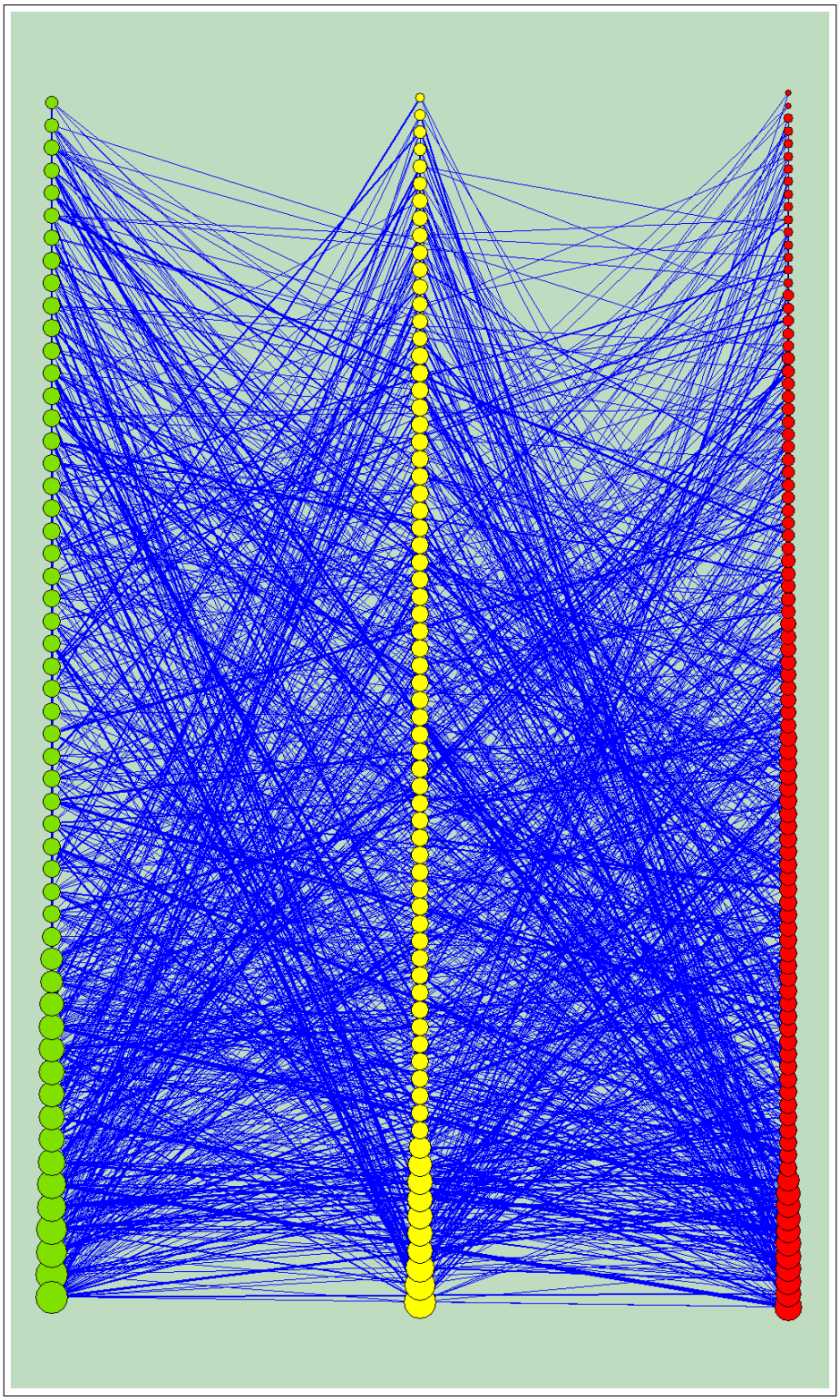


Figure 1.1: *Network structure of committees*

tournaments. These rankings are public information; Figure 1.2 shows an excerpt from the official ranking document released to participants before the 2011 tournament.[28]

The pre-tournament rankings are therefore critical public information about recent results. In the analysis that follows, we use this information as a proxy for judges' *a priori* expectations about teams' quality; for example, we define the 'favorite' in any debate as being the team with the better pre-tournament ranking.

There are two complementary reasons that these rankings work well as a proxy for judges' expectations of teams' performance. First, teams' approximate position in the rankings is well known by almost all judges. Second, even if a judge is not directly aware of teams' rankings, almost all judges know about teams' performance in recent tournaments; that is, the judges are generally aware of the underlying information on which the rankings are based. As one would expect, the difference between the rankings of two opposing teams is a significant predictor of teams' performance; as the ranking difference between two opponents narrows, the probability of the favorite winning decreases and the probability of judge disagreement increases.[29] Figure 1.3 illustrates the role of rankings for committee dissent; it shows that committee disagreement is much more common when teams are more closely ranked.

---

[28]The footnote to that document provides the ranking formula used; in order, teams are ranked by (i) the average number of wins across the Preliminary Rounds of the past three tournaments, (ii) the average number of judges won across those Preliminary Rounds, (iii) the number of wins in the most recent tournament, (iv) the number of judges won in the most recent tournament, (v) the number of wins in the second most recent tournament, (vi) the number of judges won in the second most recent tournament, (vii) the number of wins in the third most recent tournament, (viii) the number of judges won in the third most recent tournament, and (ix) alphabetically. Note that teams not having participated in the three most recent tournaments are deemed to have an 'average' of zero wins and zero judges. Note that teams are assigned to a 'group' from A to H; each team in the draw then faces one opponents from Group A, one from Group B, one from Group C, and so on. The full document is available at `http://www.schoolsdebate.com/`.

[29]To test this, we ran two probit models with 'ranking difference' as the sole explanatory variable. In the first probit, the outcome was whether the favorite wins; the estimated average marginal effect of ranking difference was about 1.5 percentage points. In the second probit, the outcome was whether the judging committee was split; the estimated average marginal effect was just under 1 percentage point. In both cases, 'ranking difference' was significant with $p < 0.001$.

WSDC 2011: PRE-TOURNAMENT RANKINGS

Rank	Country	2010		2009		2008		Average		Group
		Wins10	Judges10	Wins09	Judges09	Wins08	Judges08	WinsAvg	JudgesAvg	
1	England	8	21	8	22	8	21	**8.00**	**21.33**	**A**
2	Australia	8	23	8	21	6	18	**7.33**	**20.67**	**A**
3	Canada	8	22	7	19	7	21	**7.33**	**20.67**	**A**
4	New Zealand	6	19	8	19	8	23	**7.33**	**20.33**	**A**
5	Singapore	8	22	6	20	6	19	**6.67**	**20.33**	**A**
6	South Africa	5	15	7	20	8	22	**6.67**	**19.00**	**A**
7	Greece	8	22	6	16	6	18	**6.67**	**18.67**	**B**
8	Slovenia	7	19	6	18	6	16	**6.33**	**17.67**	**B**
9	Pakistan	5	18	6	19	7	19	**6.00**	**18.67**	**B**
10	Ireland	5	16	6	16	6	18	**5.67**	**16.67**	**B**
11	Republic of Korea	5	16	6	19	5	15	**5.33**	**16.67**	**B**
12	Wales	6	19	5	16	5	14	**5.33**	**16.33**	**B**
:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:	:	:	:
43	Japan	2	6	0	1	0	0	**0.67**	**2.33**	**H**
44	Namibia	1	6	0	2	x	x	**0.50**	**4.00**	**H**
45	Croatia	0	2	x	x	x	x	**0.00**	**2.00**	**H**
46	Barbados	x	x	x	x	x	x	**0.00**	**0.00**	**H**
47	Poland	x	x	x	x	x	x	**0.00**	**0.00**	**H**
48	Serbia	x	x	x	x	x	x	**0.00**	**0.00**	**H**

Teams are ranked by: (1) Average wins, (2) average judges, (3) wins in 2010, (4) judges in 2010, (5) wins in 2009, (6) judges in 2009, (7) wins in 2008, (8) judges in 2008, (9) alphabetically.

Figure 1.2: *Excerpt from the pre-tournament rankings, 2011*

Judges' incentives: Judges at WSDC face two primary incentives. First, every judge wants to make the 'correct' decision, by voting for the team that is more deserving of a win. There are strong norms in the international debate community — and a large degree of professional respect — for being a competent judge who accurately recognizes effective debating. Additionally, many judges pay substantial sums of money to attend the tournament, and generally take pride in participating in a high-quality tournament that is judged fairly. Second, many judges may prefer to avoid dissenting from their peers. For some judges, at least, there are strong norms that dissent is embarrassing: dissent can be seen as a strong indicator of having made the 'wrong' decision. Further, dissenting judges at WSDC often need to spend more time and effort to justify their decision, both to their peers and to the debaters.[30] Finally, dissent may arise from 'career-type' concerns; judges who are generally regarded as strong may be selected to judge in the knock-out rounds of the tournament, and tournament organizers may think less of a judge who has dissented more frequently. In an online survey conducted after the 2012 tournament, we found that a substantial share of judges admit to these kind of attitudes: 22% of respondent judges agreed that "in general, better judges are less likely to dissent", and 37% agreed that "I am more likely to worry that I have made a bad decision when I have dissented than when the result is unanimous".[31]

Of course, these two primary incentives are not limited to judges at WSDC. One would expect similar incentives for members of a typical hiring committee — where, for example, each member may have a preference over which candidate is hired, and an incentive to agree with other committee members. Similarly, one might expect judges on a court of law each to hold an opinion about the relative merits of the parties' arguments, but also a preference for judicial unanimity. In these and other contexts, estimation of the relative

[30]This point — that dissenting judges may need to spend more time to justify their decision — has also been emphasized in recent empirical work on judicial dissent: see Epstein *et al.* (2011).

[31]This online survey was described in more detail in footnote 19.

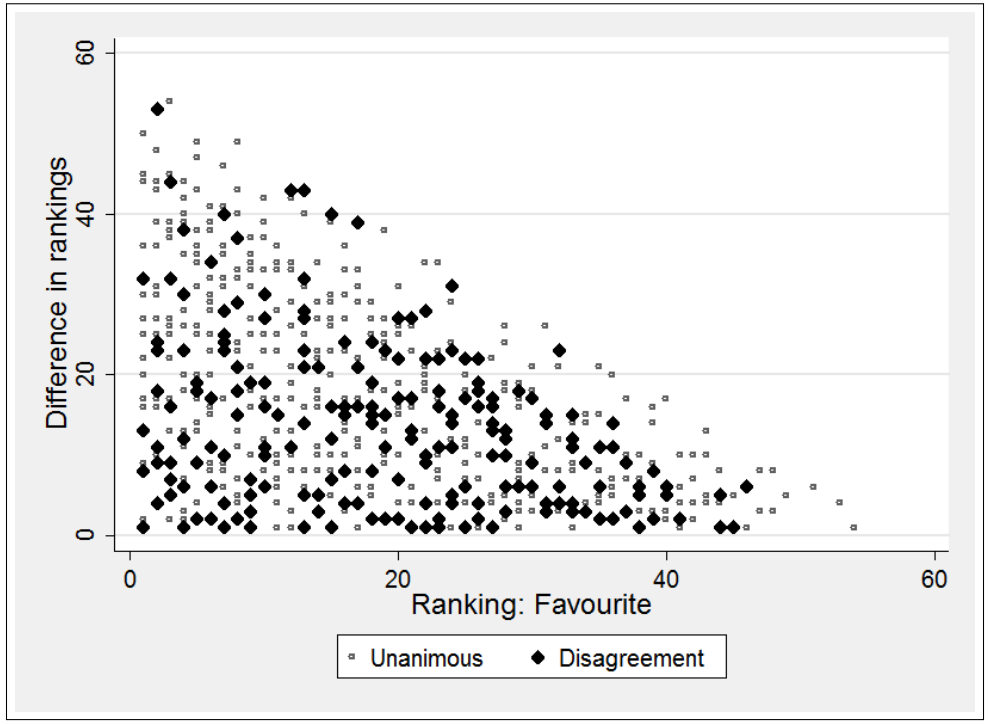


Figure 1.3: ***Dissent and pre-tournament rankings***

This figure shows the predictive value of the pre-tournament rankings. The x-axis shows the pre-tournament ranking of the favorite team; the y-axis shows the difference in pre-tournament rankings between competing teams. The figure shows that, as one would expect, committees are much significantly more likely to disagree when the competing teams' pre-tournament rankings are closer.

magnitude of these two preferences remains a pressing issue for empirical research. *How much weight do committee members place upon their individual perceptions?* Conversely, *how much do committee members value agreement with peers?* We address these questions in the remainder of the paper.

1.2.2 Identification strategy

We wish to test the relevance of previous dissent for two outcomes: (i) whether a judge votes for the favorite, and (ii) whether a judge dissents. We specify a Linear Probability Model, for judge j on committee c in round r of tournament t:

$$\begin{aligned} y_{jcrt} = \quad & \beta_1 \cdot \texttt{Dissented}_{jcrt} + \beta_2 \cdot \texttt{Dissented\_Against}_{jcrt} \\ & + \beta_3 \cdot \texttt{Distance\_Snr}_{jcrt} + \beta_4 \cdot \texttt{Distance\_Jnr}_{jcrt} \\ & + \beta_5 \cdot \texttt{Did\_Not\_Judge}_{jcrt} + \eta_{jt} + \xi_{rt} + \varepsilon_{jcrt}. \end{aligned} \tag{1.1}$$

The primary regressor of interest is $\texttt{Dissented}_{jcrt}$, a dummy for whether judge j dissented in the previous round. $\texttt{Dissented\_Against}_{jcrt}$ is a dummy for whether judge j was in a majority against a dissenter. We want to identify the effect of casting a vote in dissent separately from the effect of having merely disagreed about the relative strength of the teams; we therefore also include measures of the absolute difference in marks from judge j to his or her more 'senior' and more 'junior' peers in the previous round (respectively, $\texttt{Distance\_Snr}_{jcrt}$ and $\texttt{Distance\_Jnr}_{jcrt}$).[32] The parameters β_1, β_2, β_3 and β_4 are therefore our key parameters of interest: they measure whether judge performance is driven by dissent from a judge's peers (β_1 and β_2), distinct from 'learning' driven by distance from a peer's assessment (β_3 and β_4). Additionally, we control for whether judge j did not judge in the previous round ($\texttt{Did\_Not\_Judge}_{jcrt}$), and we allow for fixed effects for each judge in each tournament (η_{jt}) and for fixed effects for each round in each tournament (ξ_{rt}). We use

[32] Thus, for example, for a class 1 judge, $\texttt{Distance\_Snr}_{jcrt}$ records the distance to the marks of his or her class 2 peer in the previous round; $\texttt{Distance\_Jnr}_{jcrt}$ is the distance to his or her class 3 peer. 'Distance' is measured as the absolute difference in the total mark margin; for example, if the class 1 judge voted for the favorite by a margin of two, and the class 2 judge voted against the favorite by a margin of one, the absolute distance is recorded as three marks.

the two-way error structure of Cameron *et al.* (2011), clustering by committee and by judge (where we partial out η_{jt} and ξ_{rt}).

Equation 1.1 therefore allows us to test competing theories of how judges change voting behavior in response to disagreements with their peers. If judges do not care about disagreement with their peers, then neither dissenting votes nor the strength of differences in opinion should affect future judging performance: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. If judges react to having cast a dissenting vote — and, as we propose, react by placing more weight on the public signal — then we should observe that a judge who dissents is, in the following round, (i) more likely to vote for the pre-debate favorite than (s)he otherwise would be, and (ii) less likely to dissent than (s)he otherwise would be. That is, we should observe $\beta_1 > 0$ when y is a dummy for whether the judge votes for the favorite, and $\beta_1 < 0$ when y is a dummy for whether the judge dissents again (and symmetrically for β_2, if judges react to having been dissented against). If judges respond to discovering that their signal differs from their peers by updating their beliefs about the quality of their signal, then we should observe $\beta_3, \beta_4 > 0$ when y is a dummy for whether the judge votes for the favorite, and $\beta_3, \beta_4 < 0$ when y is a dummy for whether the judge dissents in the following round.

1.2.3 Regression estimates

Table 1.2 shows results for our basic specification. As predicted, we find $\beta_1 > 0$ for voting for the favorite and $\beta_1 < 0$ for dissenting. We estimate that, on average, a dissenting judge is 10 percentage points more likely to vote for the favorite (significant at the 95% confidence level), and 16 percentage points less likely to dissent (significant at the 99% confidence level). We do not find any significant effect of mark differences on probability of voting for the favorite or for dissenting.

Table 1.3 interacts the measures of dissent with dummies for each judge class, to test

Table 1.2: *Regression results: Basic specification*

	(1) **Votes for the favorite**	(2) **Dissents**
Dummy: Just dissented	0.095 (0.042)**	-0.163 (0.026)***
Dummy: Just dissented against	0.050 (0.032)	-0.015 (0.019)
Distance (senior)	-0.004 (0.003)	-0.001 (0.002)
Distance (junior)	-0.003 (0.002)	0.001 (0.002)
Dummy: Did not judge	-0.010 (0.051)	-0.037 (0.041)
Judge × tournament dummies	✓	✓
Round × tournament dummies	✓	✓
Committees	603	603
Observations	1809	1809
R^2	0.007	0.021

Standard errors in parentheses. Errors are clustered by judge and by committee.
Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

for heterogeneous effects.[33] We estimate that all three classes of judge are significantly less likely to follow one dissent with another; for all three classes, the estimated effect is about 16 percentage points. However, when we when we consider the over-weighting of public information — that is, the effect on voting for the favorite — we find a very different story. We find that the effect in the basic specification of Table 1.2 was driven wholly by class 3 judges; for that subgroup, we estimate a 22 percentage point effect of past dissent (significant at the 99% confidence level). We estimate much smaller effects — not significantly different from zero — for judges in classes A and B.[34] We either fail to observe a significant effect of mark differences on the variables of interest, or observe statistically significant but economically insignificant effects of the opposite sign to the predicted effect.

In Table 1.4, we conduct formal equality tests across the parameters of interest. We find significant heterogeneity in the probability of voting for the favorite, but not in the probability of dissenting. Together, we interpret the results of Tables 3 and 4 as showing that only class 3 judges overweight the favorite in response to having dissented.

Appendix A.2 provides further analysis of heterogeneity; it explores differences in dissent effects by judge gender, seniority of dissenting peer and tournament round.

From this evidence, we conclude that judges respond to past dissent by placing greater weight on public sources of information about the teams, and therefore become more likely to vote for the pre-debate favorite and less likely to vote in dissent. It appears that this effect is driven by a response to dissent itself, rather than being the result of judges learning about

[33]Some judges were promoted or demoted between classes during the tournament — for example, a class 2 judge who was perceived by tournament organizers to be judging well might be promoted to class 1. To avoid any potential endogeneity from such promotion/demotion, we define judge class in this specification as each judge's class at the *beginning* of a given tournament.

[34]Note that, for class 3 judges, we also find a significant effect of being dissented *against*; we estimate that class 3 judges who have just been dissented against are about 11 percentage points more likely to vote for the favorite. Our basic hypothesis does not predict such an effect, but nor does the effect run counter to such a claim.

Table 1.3: *Regression results: Heterogeneity by judge class*

	(1)	(2)	(3)	(4)	(5)	(6)
	Votes for the favorite			**Dissents**		
	Class A	**Class B**	**Class C**	**Class A**	**Class B**	**Class C**
Dummy: Just dissented	0.041	0.013	0.212	-0.183	-0.143	-0.133
	(0.073)	(0.066)	(0.075)***	(0.042)***	(0.043)***	(0.051)***
Dummy: Just dissented against	0.095	-0.040	0.106	-0.025	0.010	-0.040
	(0.051)*	(0.057)	(0.047)**	(0.026)	(0.035)	(0.041)
Distance (senior)	-0.010	0.001	-0.005	0.007	-0.004	-0.003
	(0.004)**	(0.005)	(0.005)	(0.003)**	(0.003)	(0.004)
Distance (junior)	-0.008	-0.003	0.004	0.006	0.002	-0.008
	(0.004)**	(0.004)	(0.004)	(0.002)**	(0.003)	(0.003)**
Dummy: Did not judge	-0.016	0.018	0.040	0.026	0.032	-0.142
	(0.090)	(0.121)	(0.076)	(0.077)	(0.096)	(0.059)**
Judge × tournament dummies	✓	✓	✓	✓	✓	✓
Round × tournament dummies	✓	✓	✓	✓	✓	✓
Committees	603	603	603	603	603	603
Observations	612	577	620	612	577	620
R^2	0.029	0.003	0.027	0.043	0.028	0.034

Standard errors in parentheses. Errors are clustered by judge and by committee.

'Class' refers to the judge class at the start of each tournament.

Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 1.4: ***Parameter equality tests: Heterogeneity by judge class***

	(1) **Votes for the favorite**	(2) **Dissents**
Parameter equality tests (p-values)		
H_0: Equality, 'dissented', all classes	0.104	0.692
H_0: Equality, 'dissented', class **A** *vs* **B**	0.777	0.513
H_0: Equality, 'dissented', class **A** *vs* **C**	0.097*	0.434
H_0: Equality, 'dissented', class **B** *vs* **C**	0.046**	0.878
H_0: Equality, 'dissented against', all classes	0.095*	0.593
H_0: Equality, 'dissented against', class **A** *vs* **B**	0.071*	0.414
H_0: Equality, 'dissented against', class **A** *vs* **C**	0.876	0.779
H_0: Equality, 'dissented against', class **B** *vs* **C**	0.045**	0.351

Parameter equality tests are standard Wald tests, based on the estimates (and two-way clustering) reported in Table 1.3. Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

the quality of their signal. We next use this evidence to motivate the use of a structural model to estimate the magnitude of the effect of statistical discrimination of this type.

1.3 Structural model: a three-player 'probit game'

This section introduces the theory of global games to the theory of statistical discrimination. Our earlier regression estimates show that, on average, a judge who has just dissented is substantially more likely to vote for the favorite, and less likely to dissent; as noted, both of these effects are significant. This provides clear empirical support for the hypothesis that committee members with a greater preference for coordination over-weight public information in reaching their decision — in effect, they engage in statistical discrimination. However, these regression results can say nothing about the relative magnitude of judges' preference for peer coordination against their preference for expressing their personal opinion. Without knowing this, we cannot draw conclusions beyond the present experimental context: on their own, the regression results hold little 'external validity'. For the same reason, the

regression results can tell us almost nothing about the counter-factual; that is, they can tell us nothing about the likely dynamics of a tournament in which judges do not care about committee coordination.[35]

For this, we need a structural model. Structural models are increasingly relevant for the analysis of randomized field experiments, particularly for contexts where — as here — researchers are concerned to compare the magnitude of underlying preferences, and to understand likely behavior under a counter-factual (see, for example, Duflo *et al.* (2012), Attanasio *et al.* (2012), Todd and Wolpin (2006) and Shearer (2004)).[36] At the heart of our structural model is a 'global game': a game of incomplete information, in which each player receives a signal and then acts in anticipation of the other players' choices. Global games have proved very useful for modeling coordination problems, in a variety of contexts — for example, currency crises (Morris and Shin, 1998), financial meltdowns (Allen and Morris, 1998) and political revolutions (Edmond, 2012). Global games have also been used to study committee voting — for example, by Li *et al.* (2001).

In this paper, we use a global game to formalize our intuition that committee members can engage in statistical discrimination to increase the chances of agreement; that is, debate teams' pre-tournament rankings can, like the statements of a central bank in a currency crisis,

[35]In particular, it would obviously be wrong to say, "The counter-factual is that each judge would be 10 percentage points less likely to vote for the favorite." There would be two fundamental problems with this assertion. First, voting for the favorite is a binary outcome; if a given judge has a conditional probability of voting for the favorite that is less than 10%, it is nonsensical to speak of reducing that probability by 10 percentage points. (More generally, this is a well-recognized weakness in the linearity of the Linear Probability Model: see, for example, Harrison (2011).) Second, this statement relates to the probability of a *single* judge voting for the favorite; it therefore says nothing about the probability of the favorite winning, something that inherently depends upon the behavior of the entire *committee*. For that, we would need an estimate of the correlation between the perceptions of different judges on the same committee. In our regression context, such correlation is 'controlled for' — using the two-way clustering of Cameron *et al.* (2011) — but cannot be modeled directly.

[36]Duflo *et al.* (2012, page 1265) provide a succinct and compelling justification for the use of structural models in experimental analysis, and one that applies directly to our problem: "A primary benefit of estimating a structural model of behavior is the ability to calculate outcomes under economic environments not observed in the data." See also Heckman and Smith (1995) and Orcutt and Orcutt (1968).

act as public information that plays a coordination role. In doing so, we now provide what we believe is the first application of a global game to the issue of statistical discrimination.[37] We develop a new structural model, in which a flexible assumption about the distribution of players' signals implies both a specific form for each player's optimal voting strategy and a calculable log-likelihood. Specifically, we model each player as taking a binary decision, and we model the distribution of players' signals as trivariate normal. In deference to econometric models of binary outcomes under normally-distributed error terms, we term this structure a 'three-player probit game'.[38]

The trivariate normal provides an elegant structure for allowing correlation between players' unobservable signals. In some empirical contexts, it is entirely reasonable to assume that, conditional on variables observable to the researcher, players' signals are independent: see, for example, de Paula and Tang (2012) (who study programming of radio commercial breaks) and Bajari *et al.* (2010) (on stock recommendations by equity analysts). But, in many contexts, it is unreasonable to assume that the common elements to players' signals are observed by the researcher. The present empirical context provides one illustration: each committee watches the same debate, so receives correlated signals (in the form of speakers' presentations), and those signals cannot be captured fully by any variables observed by the researcher.[39] The trivariate normal implies a very convenient form for each player's best response function, as well as a calculable log-likelihood — and does so while allowing for

[37]For example, Lang and Lehman (2012) provide an excellent review of recent results on racial discrimination in labor markets, including statistical discrimination; that review does not include any discussion of committee effects.

[38]One way of thinking about our model is that it provides microfoundations for the trivariate probit. The trivariate probit is a standard method for estimating the determinants of correlated binary outcomes, where those binary outcomes are grouped into triples. However, on its own, the trivariate probit says nothing about the incentive structures that face a group of three decision-makers each taking a binary choice. Our model shows how a particular version of the trivariate probit can be given coherent choice-theoretic/game-theoretic foundations. We limit attention to the three-player case for simplicity; because all of our committees comprise three judges, nothing would be gained in our empirical application by considering more players. But all of the results here could extend to higher dimensions — albeit with inevitable additional computational complexity.

[39]Our structural estimates, reported shortly, support this claim; we strongly reject a null hypothesis that signals are conditionally independent.

correlated player signals. To avoid resting our identification solely upon a distributional assumption, we exploit the number of previous dissents as an excludable shock to player payoffs.[40]

1.3.1 Committee voting as a global game

Model setup

We model each committee as an independent Bayesian game between three players. For each committee, we denote judge class by $i \in \{1, 2, 3\}$. Each judge i receives a signal x_i, and then chooses whether to vote for the favorite ($a_i = 1$) or against ($a_i = 0$). Judge i receives utility from two mechanisms: (i) from voting for the team that (s)he prefers (where the strength of that preference is determined by the signal x_i),[41] and (ii) from agreeing with judge j and/or with judge k. We treat these mechanisms as additively separable.[42] Note that, for example, δ_{ij} measures the utility gain for judge i from voting with judge j, and that δ_i measures the gain from agreeing with *both* judges k and j.

$$U_i(a_i; a_j, a_k, x_i) = \begin{cases} x_i + \delta_i & \text{if } a_i = 1, a_j = 1, a_k = 1; \\ x_i + \delta_{ij} & \text{if } a_i = 1, a_j = 1, a_k = 0; \\ x_i + \delta_{ik} & \text{if } a_i = 1, a_j = 0, a_k = 1; \\ x_i & \text{if } a_i = 1, a_j = 0, a_k = 0; \\ 0 & \text{if } a_i = 0, a_j = 1, a_k = 1; \\ \delta_{ij} & \text{if } a_i = 0, a_j = 1, a_k = 0; \\ \delta_{ik} & \text{if } a_i = 0, a_j = 0, a_k = 1; \\ \delta_i & \text{if } a_i = 0, a_j = 0, a_k = 0. \end{cases} \tag{1.2}$$

[40] In this sense, we take a similar approach to Grieco (2011), who uses excludable covariates and a bivariate normal distribution to identify a binary choice game between two players.

[41] In taking this approach, our model differs from the standard approach to modeling committee decision-making in two ways: by assuming that utility is related to a private value, rather than a common fundamental, and by assuming that judges care about their *vote*, rather than the decision of the entire committee (see, for example, Austen-Smith and Banks (1996) and Feddersen and Pesendorfer (1996)). Our approach allows us to focus on the coordination mechanism and abstract away from other reasons for judges to give weight to a public signal, and avoids the difficulties associated with judge behavior depending on whether or not their vote is pivotal. Loosening either assumption does not significantly change the theoretical results, while complicating the estimation. Note also that our approach allows for the possibility of cases with lower and higher stakes in judge decision-making, in contrast to the standard approach of assuming a fixed cost associated with Type-I and Type-II errors.

[42] This kind of additively separable 'reduced form' specification is standard for structural models of incomplete information: see, for example, de Paula and Tang (2012), Bajari *et al.* (2010) and Grieco (2011).

We assume that each judge weakly prefers agreement over dissent, and agreeing with both peers over agreeing with just one: $\delta_i \geq \delta_{ij}, \delta_{ik} \geq 0$.

For each committee, the distribution of signals is trivariate normal (where we assume positive correlations, $\rho_{12}, \rho_{13}, \rho_{23} > 0$):[43,44]

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \right). \tag{1.3}$$

The signal x_i therefore plays a dual role: it directly affects the relative utility of voting for the favorite, and it determines the conditional expectation of the other judges' signals:

$$\begin{pmatrix} x_j \\ x_k \end{pmatrix} \Bigg| \; x_i \sim \mathcal{N}\left[\begin{pmatrix} \mu_j + \rho_{ij} \cdot (x_i - \mu_i) \\ \mu_k + \rho_{ik} \cdot (x_i - \mu_i) \end{pmatrix}, \begin{pmatrix} 1 - \rho_{ij}^2 & \rho_{jk} - \rho_{ij} \cdot \rho_{ik} \\ \rho_{jk} - \rho_{ij} \cdot \rho_{ik} & 1 - \rho_{ik}^2 \end{pmatrix} \right]. \tag{1.4}$$

Judge i must choose a best response $a_i^*(x_i)$; we limit attention to cutoff strategies: $a_i^*(x_i) = \mathbf{1}(x_i \geq x_i^*)$, where x_i^* denotes the cutoff and $\mathbf{1}(\cdot)$ the indicator function.[45] Judge i must be indifferent between $a_i = 0$ and $a_i = 1$ if $x_i = x_i^*$; that is,

$$\begin{aligned} x_i^* = &\left[\Pr\left(a_j = 0, a_k = 0 \mid x_i = x_i^*\right) - \Pr\left(a_j = 1, a_k = 1 \mid x_i = x_i^*\right) \right] \cdot \delta_i + \\ &\left[\Pr\left(a_j = 0, a_k = 1 \mid x_i = x_i^*\right) - \Pr\left(a_j = 1, a_k = 0 \mid x_i = x_i^*\right) \right] \cdot \left(\delta_{ij} - \delta_{ik}\right). \end{aligned} \tag{1.5}$$

[43] Readers who are not used to seeing global games model expressed with correlated signals, rather than an underlying fundamental, should see Morris and Shin (2006).

[44] Of course, as with any trivariate normal, we must also assume a positive definite covariance matrix; this implies the further restrictions $\rho_{12}, \rho_{13}, \rho_{23} < 1$ and $1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12} \cdot \rho_{13} \cdot \rho_{23} > 0$. Note that the restriction $\text{Var}(x_1) = \text{Var}(x_2) = \text{Var}(x_3) = 1$ is made without loss of generality; if we were to parameterize these variances, all of our results would simply rescale by those new parameters. This is the same reasoning that justifies the same restriction for the trivariate probit.

[45] Focusing attention on cutoff strategies is common in the global games literature (see, for example, Morris and Shin (2003)). Iterated elimination of dominated strategies implies that where a unique equilibrium exists in these games, it will be monotone.

Equilibrium characterization

An equilibrium is then defined by a vector of cutoffs $\{x_i^*, x_j^*, x_k^*\}$ such that each player is indifferent between $a_i = 0$ and $a_i = 1$ given the expected play of the other players. With the information structure described above, x_i^* is therefore defined by:

$$0 = x_i^* + \left\{ \Phi_2 \left[-\alpha_j(x_i^*), -\alpha_k(x_i^*), \omega_{jk} \right] - \Phi_2 \left[\alpha_j(x_i^*), \alpha_k(x_i^*), \omega_{jk} \right] \right\} \delta_i$$
$$+ \left\{ \Phi_2 \left[-\alpha_j(x_i^*), \alpha_k(x_i^*), -\omega_{jk} \right] - \Phi_2 \left[\alpha_j(x_i^*), -\alpha_k(x_i^*), -\omega_{jk} \right] \right\} (\delta_{ij} - \delta_{ik})$$

where:

$$\alpha_j(x_i^*) = \frac{x_j^* - \mu_j - \rho_{ij}(x_i^* - \mu_i)}{\sqrt{1 - \rho_{ij}^2}},$$
$$\alpha_k(x_i^*) = \frac{x_k^* - \mu_k - \rho_{ik}(x_i^* - \mu_i)}{\sqrt{1 - \rho_{ik}^2}}, \text{and}$$
$$\omega_{jk} = \frac{\rho_{jk} - \rho_{ij} \cdot \rho_{ik}}{\sqrt{\left(1 - \rho_{ij}^2\right) \cdot \left(1 - \rho_{ik}^2\right)}}.$$

x_j^* and x_k^* are defined analogously.

Proposition 1 (conditional state monotonicity) *For judge i, the difference in utility between $a_i = 1$ and $a_i = 0$ is monotonically increasing in x_i, and therefore each judge has a unique cutoff x_i^*.*

Proof: Proofs are in Appendix A.1.

Remark. It is worth noting that while it is sufficient for Proposition 1 that no judges have a preference for 'discoordination' with another judge ($\delta_i \geq \delta_{ij}, \delta_{ik}$), it is not necessary.[46] One can imagine some circumstances in which one committee members may have a preference

[46]Further details are available in the proof.

for discoordination with another member — for example, judges on appellate courts are sometimes portrayed as holding such preferences — but we argue that imposing strategic complementarities is entirely reasonable for the vast majority of committee contexts, including our field experiment.

Proposition 2 (unique equilibrium) *It is sufficient for the existence of a unique equilibrium that, for each judge i,*

$$\delta_i < \sqrt{\frac{\pi}{2(1-\omega_{jk}^2)}} \cdot \left(\sqrt{\frac{1-\rho_{ij}}{1+\rho_{ij}}} + \sqrt{\frac{1-\rho_{ik}}{1+\rho_{ik}}} \right)^{-1} \tag{1.6}$$

$$\tag{1.7}$$

Remark. Proposition 2 is an extension to three players of the results in Morris and Shin (2006). However, we have fixed the noise of agents' signals to unity and allowed the returns to coordination to vary, while Morris and Shin (2006) (and most of the global games literature) fix the returns to coordination and allowed the noise of agents' signals to vary. Put in this way, Proposition 2 generates a new insight into uniqueness conditions for global games: uniqueness requires that players do not care *too much* about coordination. If the returns to coordination are too high relative to the correlation of agents' signals, then we cannot guarantee that there do not exist multiple equilibria in which players coordinate on selecting either the favorite or the underdog for regions in which they do not have a dominant strategy.

It is also worth noting that the shift to three players generates bounds that are tighter than the equivalent restrictions for two players. This is because the proof of uniqueness relies on translations of monotone strategies, and the same shift in i's signal with three players generates a greater potential change in payoffs than it would if i had only one other player to coordinate with. This is the case even when one of ρ_{ij} or ρ_{ik} are equal to zero, and i's signal is completely uninformative of another player's.

Interpretation and comparative statics

Consider a "no-coordination" benchmark by setting each of the δ terms equal to zero. Then each judge selects $x_i^* = 0$, and only votes for the favorite if he or she believes that the favorite genuinely won the debate and will receive private utility from voting for them.

Now suppose that — in a way we make precise shortly — all three judges broadly agree on who the expected winner of the debate is and by how much. In that case, a preference for coordination generates the statistical discrimination that we are testing: $\delta_i > 0$ implies $x_i^* < 0$. Judge i therefore has a range of signals $x_i^* < x_i < 0$ where he or she would privately prefer to vote for the underdog, but instead votes for the favorite out of a desire to increase their chances of coordinating with their fellow judges.

Proposition 3 (Statistical Discrimination) *If $\mu_m - \rho_{mn}\mu_n > 0 \; \forall \; m, n \in i, j, k$, then $x_i^* \leq 0 \; \forall \; i$, with strict inequality if $\delta_i > 0$.*

Note that if $\mu_i = \mu_j = \mu_k$, then the required condition in Proposition 3 holds straightforwardly, and so any desire for coordination leads to statistical discrimination. The reason for the required condition is that it guarantees that any judge does not have a signal distribution with a mean so far away from the other judges, and therefore a much higher ex ante probability of voting for the favorite, that increasing the rewards to coordination leads that judge to become *less* likely to vote for the favorite. In this scenario, incentives to coordinate can play the role we might otherwise have thought they would play - constraining the behavior of extremist judges. Note, however, that even in that case incentives to coordinate still generate statistical discrimination in the other judges.

The key result we require in this section to link the model to the empirical exercise is a comparative static on x_i^* with respect to each of the δ terms.

Proposition 4 (Comparative Statics) *The comparative statics with respect to the δ parameters are monotonic and of the following form:*

$$sign(\frac{dx_i^*}{d\delta_i}) = sign(\frac{\partial x_i^*}{\partial \delta_i}) = -sign\left(\Phi_2\left[-\alpha_j(x_i), -\alpha_k, \omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i), \alpha_k(x_i), \omega_{jk}\right]\right)$$

$$sign(\frac{dx_i^*}{d\delta_{ij}}) = sign(\frac{\partial x_i^*}{\partial \delta_{ij}}) = -sign\left(\Phi_2\left[-\alpha_j(x_i), \alpha_k, \omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i), -\alpha_k(x_i), \omega_{jk}\right]\right)$$

$$sign(\frac{dx_i^*}{d\delta_{ik}}) = sign(\frac{\partial x_i^*}{\partial \delta_{ik}}) = -sign\left(\Phi_2\left[\alpha_j(x_i), -\alpha_k, \omega_{jk}\right] - \Phi_2\left[-\alpha_j(x_i), \alpha_k(x_i), \omega_{jk}\right]\right)$$

If $\delta_{ij} = \delta_{ik}$ and $\mu_m - \rho_{mn}\mu_n > 0 \; \forall \; m, n \in i, j, k$, then x_i is monotonically decreasing in δ_i.

Remark. For the purposes of model identification, it is sufficient that these effects be monotone (see Proposition 5). However, Proposition 4 also demonstrates a further point of independent interest. If the required condition in Proposition 3 holds and $\delta_{ij} = \delta_{ik}$, then each of the comparative statics in Proposition 4 is negative: that is, if we have the conditions for statistical discrimination and judges care equally about their peers, then the comparative statics with respect to the incentives to coordinate will be such that greater incentives to coordinate generate greater statistical discrimination.[47]

If, however, the required condition in Proposition 3 does not hold, then the comparative statics of judge i's cutoffs with respect to their incentives to coordinate are still monotone holding the other judge's coordination parameters constant, but the sign of the relationship depends on the values of i's other δ parameters and the δ parameters for j and k. The possibility of of this effect changing sign is not as far-fetched as it may sound: in our structural estimation, the Proposition 3 condition holds for most ranking differences, but not for the very highest levels.[48] While we should be careful not to place too much weight

[47] If judges care differently about their peers ($\delta_{ij} \neq \delta_{ik}$), then this effect will likely still occur, but we cannot guarantee it. Our parameterization below imposes $\delta_{ij} = \delta_{ik}$.

[48] Using our estimates from Table 1.5 and the greatest ranking difference in our data (56), $\mu_1 - \rho_{12}\mu_2 = -0.219$, violating the condition in Proposition 3. Note, however, that we also estimate $\delta_1 = \delta_2 = 0$, and so past dissent appears to have no effect on the behavior of type 1 and 2 judges.

on data outliers, this suggests that there are potentially some debates in which greater incentives to coordinate can lead individual judges to view the underdog more favorably.

1.3.2 Structural implementation

Parameterization: We estimate common values for ρ_{12}, ρ_{13} and ρ_{23} across all committees. For each committee c, we denote the difference in pre-tournament rankings by $R_c > 0$. We allow this ranking difference to shift judges' signal means; for flexibility, we adopt a quadratic specification, and allow a different relationship for each judge class:[49]

$$\mu_{1c} = \beta_1 \cdot R_c + \gamma_1 \cdot R_c^2; \tag{1.8}$$

$$\mu_{2c} = \beta_2 \cdot R_c + \gamma_2 \cdot R_c^2; \tag{1.9}$$

$$\mu_{3c} = \beta_3 \cdot R_c + \gamma_3 \cdot R_c^2. \tag{1.10}$$

We use the dummy D_{ic} to denote that judge i on committee c dissented in the previous round.[50] We allow this dummy to shift each judge's preference for agreement. We allow different classes of judges to be differentially affected by previous dissent, and we impose that each judge is indifferent between agreeing with one peer and agreeing with two:

$$\delta_{1c} = \delta_{12c} = \delta_{13c} = \delta_1 \cdot D_{1c}; \tag{1.11}$$

$$\delta_{2c} = \delta_{21c} = \delta_{23c} = \delta_2 \cdot D_{2c}; \tag{1.12}$$

$$\delta_{3c} = \delta_{31c} = \delta_{32c} = \delta_3 \cdot D_{3c}. \tag{1.13}$$

Equations 1.11 – 1.13 reflect the central intuition emerging from the reduced-form estimates: that past dissent increases a judge's preference for coordination. The equations also show two important limitations of this estimation method. First, the current experimental context does not allow us to identify δ_i, δ_{ij} and δ_{ik} separately; this is because the exogenous

[49] Equations 1.8 – 1.10 imply that, in the hypothetical case that two teams were equally matched ($R_c = 0$), then $\mu_{c1} = \mu_{c2} = \mu_{c3} = 0$. This is exactly as we would expect and require.

[50] That is, D_{ic} is equivalent to $\texttt{Dissented}_{jcrt}$ in the reduced-form estimation.

variation (past dissent) operates at the level of the individual judge. We *could* use the present structural methodology to separately identify δ_i, δ_{ij} and δ_{ik}, if we observed some exogenous shock operating at the level of the *relationship* between judges i and j.[51] Second, we use the structural model to identify the preference for coordination driven by past dissent, but we do not seek to identify the preference for coordination generally. That is, if $D_{ic} = 0$, we impose $\delta_{ic} = 0$.[52]

Constraints: We constrain the estimation so that $\rho_{12}, \rho_{13}, \rho_{23} \in [0.01, 0.99]$, and so that the covariance matrix is positive definite.[53] We impose $\delta_1, \delta_2, \delta_3 \geq 0$. Together, these constraints ensure conditional state monotonicity (Proposition 1). Additionally, we impose the single-equilibrium condition of Proposition 2.[54]

Identification:

Proposition 5 (global identification of the three-player probit game) *Assume that the conditions in Proposition 1 and Proposition 2 hold, and that R_c takes at least two unique values. Then the structural model is globally identified.*

Estimation method: The proof of identification relies only upon a subset of cases on (D_{1c}, D_{2c}, D_{3c}). But, to estimate efficiently, we use all of our data. Specifically, we use

[51]This might be more realistic in an applied IO context; if this were a coordination game between three firms, for example, one might imagine an exogenous regulatory shock changing the coordination incentive between just two of those firms. If such variation were observed, our global identification result could readily be extended to identify $(\delta_i, \delta_{ij}, \delta_{ik})$ generally; details available on request.

[52]We leave for future work the question of whether δ_{ic} can be separately identified for judges who have not just dissented — for example, by exploiting judges' responses to the past dissent of their peers (see, by analogy, Grieco (2011)). The normalization appears reasonable in our empirical context: when we relax the normalization, we obtain $\ell = -843.793$, implying $p = 0.612$ on the hypothesis $H_0 : \delta_{1c} = \delta_{2c} = \delta_{3c} = 0$ when $D_{1c} = D_{2c} = D_{3c} = 0$.

[53]That is, we additionally impose $1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12} \cdot \rho_{13} \cdot \rho_{23} > 0$.

[54]We impose uniqueness for simplicity; we find it completely implausible in this particular context that judges could care so little about their perception of the debate that multiple equilibria could exist. But this restriction could be relaxed for a different empirical context; for example, by using an equilibrium selection rule, or by allowing equilibria to co-exist according to a discrete finite mixture distribution. See de Paula and Tang (2012) and Su (2012).

Maximum Likelihood with a nested fixed-point approach. Denote the stacked parameter vector as $\boldsymbol{\theta}$. Then, for some candidate value for $\boldsymbol{\theta}$, the inner loop solves the game for each committee c in the dataset,[55] given the ranking data: $(x_{1c}^*(\boldsymbol{\theta}; R_c), x_{2c}^*(\boldsymbol{\theta}; R_c), x_{3c}^*(\boldsymbol{\theta}; R_c))$. We then calculate the log-likelihood $\ell(\boldsymbol{\theta})$ using a standard triprobit structure (where we approximate the *cdf* of the trivariate normal using the method of Genz (2004)).[56] The outer loop updates using a Sequential Quadratic Program. We calculate p-values for our parameter estimates using Likelihood Ratio tests.

1.3.3 Structural estimates

Table 1.5 reports the resulting structural estimates. Several points deserve noting. First, all of the covariance terms (ρ_{12}, ρ_{13} and ρ_{23}) are significantly different from zero (indeed, the p-value for each the covariance terms is less than 1^{-12}). This provides strong empirical support for our refusal to assume that signals are conditionally independent. That assumption may be reasonable in other contexts, but is clearly not reasonable in this context, where players react to their perceptions of a common event (that is, to the debate they observe).[57] Second, the covariance terms are significantly different from each other; when we use a Likelihood Ratio test for the null hypothesis $\rho_{12} = \rho_{13} = \rho_{23}$, we obtain $p < 0.004$. We find that, as one might expect, the signals received by class 1 and class 2 judges are more similar to each other than either is to the signals received by class 3 judges (that is, $\rho_{12} > \rho_{13} \approx \rho_{23}$).

---

[55] Note that this aspect of the problem — and the calculation of the log-likelihood — is trivially parallelizable. We exploit this to improve dramatically the speed of estimation.

[56] Formally, we calculate the log-likelihood for committee c as:

$$\begin{aligned}\ell_c\left(\boldsymbol{\theta}; a_{1c}, a_{2c}, a_{3c}\right) = \ln \Phi_3 \Big[& (2a_{1c}-1) \cdot \left(\beta_1 \cdot R_c + \gamma_1 \cdot R_c^2 - x_1^*(\boldsymbol{\theta}; R_c)\right), \\ & (2a_{2c}-1) \cdot \left(\beta_2 \cdot R_c + \gamma_2 \cdot R_c^2 - x_2^*(\boldsymbol{\theta}; R_c)\right) \\ & (2a_{3c}-1) \cdot \left(\beta_3 \cdot R_c + \gamma_3 \cdot R_c^2 - x_3^*(\boldsymbol{\theta}; R_c)\right), \rho_{12}, \rho_{23}, \rho_{13} \Big],\end{aligned}$$

where $\Phi_3(\cdot)$ refers to the *cdf* of the standard trivariate normal. We treat draws of (x_1, x_2, x_3) as independent across committees, so the sample log-likelihood is $\ell(\boldsymbol{\theta}) = \sum_{c=1}^{603} \ell_c(\boldsymbol{\theta})$.

[57] Iaryczower *et al.* (2013) consider a similar empirical context, in which appellate judges observe arguments between competing parties in court cases.

Table 1.5: ***Structural estimates: Basic specification***

parameter	**estimate**	ℓ_r	LR	p-value
ρ_{12}	**0.725**	-902.433	115.467	0.000^{***}
ρ_{13}	**0.526**	-871.012	52.625	0.000^{***}
ρ_{23}	**0.542**	-873.090	56.781	0.000^{***}
β_1	**0.064**	-869.736	50.074	0.000^{***}
γ_1	**-6.912e**$^{-4}$	-847.440	5.480	0.019^{**}
β_2	**0.045**	-855.134	20.868	0.000^{***}
γ_2	**-8.411e**$^{-5}$	-844.735	0.071	0.791
β_3	**0.035**	-852.373	15.346	0.000^{***}
γ_3	**4.480e**$^{-5}$	-844.712	0.025	0.875
δ_1	**0.000**	-844.699	0.000	1.000
δ_2	**0.000**	-844.699	0.000	1.000
δ_3	**0.871**	-845.427	1.455	0.228
LOG-LIKELIHOOD (ℓ_u)		-844.699		

Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
We calculate p-values using the the χ^2 distribution with one degree of freedom. Note that, for the covariance terms, we test $H_0 : \rho_{12} = 0.01$, $H_0 : \rho_{13} = 0.01$ and $H_0 : \rho_{23} = 0.01$; we take this approach because we use 0.01 as the lower bound on the covariance terms throughout.

Most importantly, we find that past dissent generates a substantial desire for coordination, but that this is limited to class 3 judges (that is, we estimate $\delta_1 = 0$, $\delta_2 = 0$ and $\delta_3 = 0.871$, with $p = 0.228$ for $H_0 : \delta_3 = 0$). We find that dissent aversion is limited to class 3 judges, but that those judges have a relatively large preference for agreement; this reflects the reduced-form estimation results in Table 1.3. There are several complementary interpretations of the estimated parameter for δ_3. We first calculate the equivalent variation — the proportion of decisions that a class 3 judge would be willing to change under full information in order to avert dissent.[58] This is shown in Figure 1.4. The equivalent variation is large; for the median ranking difference of 15, we estimate that class 3 judges would, under full information, shift about 22% of their decisions to avoid dissent.

Figure 1.4 shows the equivalent variation under an illustrative hypothetical of full information; it illustrates the utility cost of dissent, but does not consider equilibrium behavior in the Bayesian game. To consider equilibrium behavior, we calculate the probability that — in equilibrium — a class 3 judge votes for the favorite, conditional on just having dissented.[59] In Figure 1.5, we show how this probability changes between $\delta_3 = 0.871$ and $\delta_3 = 0$. Again, the difference is substantial: at a ranking difference of 15, class 3 judges are about eight percentage points more likely to vote for the favorite as a result of their preference for coordination.

Finally, in Figure 1.6, we show the effect of dissent aversion on the probability that the favorite wins, conditional on the class 3 judge having just dissented. We calculate the probability that the favorite wins (either in a split decision or a unanimous decision), and show how this probability would change if $\delta_3 = 0$. We find only a very small effect: at

[58] This can be calculated straightforwardly. From equations 1.3.1 and 1.10, a class 3 judge would be indifferent, under full information, between using a cutoff of 0 and dissenting or using a cutoff of $-\delta_3$ and not dissenting. For a given ranking difference R_c, a class 3 judge would therefore be willing to change a proportion of decisions $\Phi\left(\beta_3 \cdot R_c + \gamma_3 \cdot R_c^2 + \delta_3\right) - \Phi\left(\beta_3 \cdot R_c + \gamma_3 \cdot R_c^2\right)$ to avoid dissent.

[59] This requires solving the game numerically, just as we do for the nested fixed-point estimation algorithm.

a ranking difference of 15, the favorite's probability of winning increases by only about two percentage points. This is not surprising, given our earlier estimates of the covariance parameters: because class 1 and class 2 judges receive quite similar signals ($\hat{\rho} = 0.725$), it is relatively unusual that the class 3 judge is a pivotal voter.

In sum, we draw two general conclusions. First, committee members' preference for coordination can be large, and this can lead committee members to over-weight publication information, thus generating statistical discrimination. Second, the effect of such discrimination can only be understood by estimating the covariance between committee members' signals; in this context, we find that the overall effect of the desire for coordination is negligible, even though effect on the decisions of the least experienced judges can be quite large. Together, these conclusions illustrate the value of combining experimental and structural methods to understand committee voting behavior.

1.3.4 Model validation through hypothesis registration

Model validation is an important challenge for structural analysis.[60] In this paper, we take a novel approach to structural model validation: we used our structural model to make testable predictions about the 2013 World Schools Debating Championships, and registered the model and predictions at the J-PAL Hypothesis Registry.[61] To our knowledge, this approach to out-of-sample validation of a structural model has not been used before.

Table 1.6 summarizes both in-sample and out-of-sample predictions; the out-of-sample predictions (the final three columns) relate directly to the registered hypotheses. In each

[60] See, for example, Keane (2010, page 18), who argues, "It has often been treated as a feat worthy of praise to simply estimate a structural model, regardless of whether the model can be shown to provide a good fit to the data, or perform well in out-of-sample predictive exercises. I see no reason why an estimated structural model should move my priors about, say, the likely impact of a policy intervention, if it fits the in-sample data poorly and/or has not been shown to perform reasonably well in any validation exercises. Structural econometricians need to do a much better job in this area in order for structural work to gain wider acceptance."

[61] `http://www.povertyactionlab.org/Hypothesis-Registry`.

case, we took the particular tournament structure (*i.e.* the assigned match-ups between opponents of differing pre-tournament rankings), and ran 1000 simulated versions of the tournament. We then compare moments between actual and simulated tournaments. In each case, we compare the observed moment to the mean of the simulated distribution; we also report the corresponding percentile of the simulated distribution.[62] We underline measured moments lying outside a 90% confidence interval in our simulated distribution (that is, values whose percentile is less than 5% or greater than 95%).

In general, the model performs well, both in-sample and out-of-sample. Of the 72 moments reported, nine lie outside the 90% confidence interval from the simulated data; of the 24 moment predictions made for the 2013 tournament, three lie outside the same interval. Insofar as the model performs poorly, it does so in predicting the probability that Class 2 dissents. This appears to be driven by an unusual variation in Class 2 behavior between tournaments; something that lies beyond the scope of our model framework.

1.4 Conclusions

The key result of this paper is to demonstrate that incentives to coordinate in committees can cause statistical discrimination. This proposed mechanism is backed up by a field experiment that gave results consistent with its operation, a simple coordination model that generates the proposed effect (depending on parameter values) and a structural estimation of that model that suggests that the effect is active, but is not having a large effect on the outcome of the environment we study. We further studied the external validity of this model by making out-of-sample predictions and registering them with J-PAL.

The next step in this research agenda is to compare these static results with a fully

[62] Our registered document contained graphs of the empirical CDFs from our simulation exercise; in this way, we registered not merely a predicted mean of each moment, but its entire distribution.

Table 1.6: *Goodness of fit: In-sample and out-of-sample structural predictions*

	in-sample									out-of-sample		
	2010			2011			2012			2013		
moment (PROBABILITY)	*pred.* (%)	*actual* (%)	*perc.* (%)	*pred.* (%)	*actual* (%)	*perc.* (%)	*pred.* (%)	*actual* (%)	*perc.* (%)	*pred.* (%)	*actual* (%)	*perc.* (%)
Class 1												
judge dissents	8.8	8.9	**59.4**	9.0	8.9	**53.7**	8.9	8.4	**45.1**	8.8	10.7	**82.3**
... if just dissented	8.6	7.7	**50.8**	8.5	0.0	**31.2**	8.5	8.3	**52.4**	8.4	0.0	**32.3**
... if not just dissented	8.8	8.9	**54.3**	9.0	9.3	**57.5**	9.0	8.4	**41.3**	8.8	11.8	**90.9**
judge votes for the favorite	76.4	76.0	**39.4**	75.6	76.6	**64.3**	75.8	78.5	**81.3**	76.3	74.3	**22.6**
... if just dissented	76.3	46.2	**1.1**	75.1	90.0	**92.0**	75.7	66.7	**24.1**	76.1	76.5	**48.0**
... if not just dissented	76.4	78.2	**71.4**	75.6	75.8	**51.8**	75.8	79.3	**88.2**	76.3	74.1	**21.6**
Class 2												
judge dissents	8.4	4.7	**3.6**	8.6	9.4	**68.3**	8.7	12.0	**95.3**	8.5	13.4	**98.3**
... if just dissented	7.7	16.7	**89.7**	8.4	7.1	**44.9**	7.8	14.3	**83.4**	8.2	10.0	**65.5**
... if not just dissented	8.5	3.9	**0.6**	8.6	9.6	**66.2**	8.7	11.8	**90.4**	8.6	13.8	**98.9**
judge votes for the favorite	74.7	75.0	**56.4**	73.6	72.9	**37.5**	73.7	71.7	**24.5**	74.2	75.9	**69.7**
... if just dissented	74.9	75.0	**50.5**	73.6	57.1	**8.3**	74.0	71.4	**36.4**	74.4	75.0	**51.7**
... if not just dissented	74.6	75.0	**53.5**	73.6	74.2	**57.3**	73.7	71.8	**25.6**	74.2	76.1	**72.1**
Class 3												
judge dissents	15.0	18.2	**89.3**	15.4	15.6	**51.8**	15.4	12.6	**13.3**	15.2	15.5	**53.0**
... if just dissented	12.6	5.0	**15.6**	13.2	14.3	**59.7**	13.1	26.7	**97.7**	13.1	19.1	**82.1**
... if not just dissented	15.4	19.8	**94.3**	15.7	15.8	**53.7**	15.8	11.4	**4.2**	15.5	15.1	**44.8**
judge votes for the favorite	71.9	71.9	**51.9**	70.8	72.9	**71.3**	70.9	67.0	**11.4**	71.5	71.7	**50.1**
... if just dissented	77.0	70.0	**21.5**	75.7	66.7	**16.9**	75.8	80.0	**68.4**	76.4	57.1	**2.1**
... if not just dissented	71.1	72.1	**58.4**	70.1	73.7	**85.5**	70.2	65.9	**11.8**	70.7	73.5	**76.6**

This table reports in-sample and out-of-sample predictions for the 2010, 2011, 2012 and 2013 World Schools Debating Championships. Predictions are formed from 1000 simulations, using the structural estimates from Table 1.5. 'Pred' reports the mean predicted statistic. 'Actual' reports the actual statistic. 'Perc' reports the percentile for the actual statistic in the empirical distribution of 1000 simulations. 'Judge dissents' reports the unconditional probability of a judge dissenting. 'Judge dissents... if just dissented' reports the probability of a judge dissenting, conditional on having just dissented in the previous round. 'Judge dissents... if not just dissented' reports the probability of a judge dissenting, conditional on not having just dissented in the previous round. The same structure applies for statistics reporting 'votes for the favorite'.

dynamic model. Since the World Schools Debating Championships is of finite length, judge actions in a dynamic model are solvable by backward induction, and it is possible to develop equivalent identification strategies to the ones proposed in this paper (see Appendix A.4). It would then be possible then test to see if judges are become more dissent averse as a behavioral response to past dissent, or if they are optimizing over the course of the tournament. Beyond these specific circumstances, this has implications for models of committee decision-making where dynamic aspects are a more natural concern (such as judicial decision-making).

The other set of interesting questions relate to the policy implications of this mechanism: what new trade-offs does this generate for optimal committee composition? This result likely generates a new argument leaning against the Condorcet jury theorem, alongside the now familiar argument of free-riding in information acquisition. In settings of endogenous information acquisition, the possibility of dissent aversion might also generate an additional novel mechanism: information acquisition decisions could potentially be strategic substitutes, in that if a fellow judge invests in information, it is in your interest to invest as well to assist in coordination. To investigate these questions, it is necessary to reintroduce both common fundamentals and preferences over committee outcomes, rather than expressive voting; we are currently developing this approach in a separate paper.

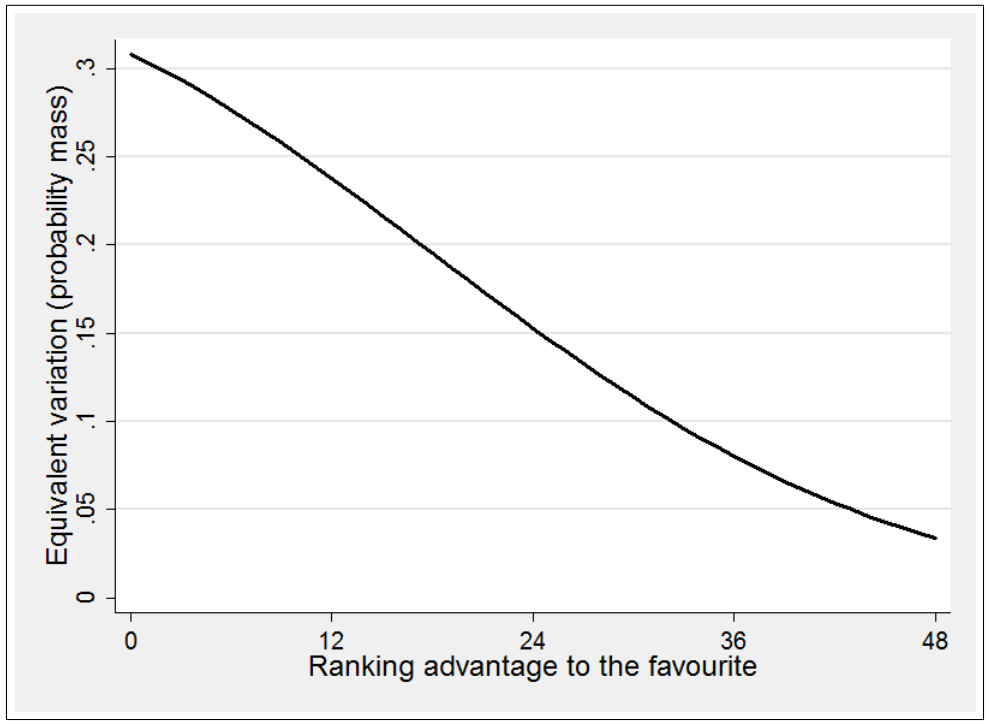


Figure 1.4: ***Proportion of decisions that a class 3 judge would shift to avoid dissent***

This figure shows the proportion of decisions that a class 3 judge would change, under full information, to avoid dissent: our measure of equivalent variation in this committee context. We calculate this as $\Phi\left(\beta_3 \cdot R_c + \gamma_3 \cdot R_c^2 + \delta_3\right) - \Phi\left(\beta_3 \cdot R_c + \gamma_3 \cdot R_c^2\right)$.

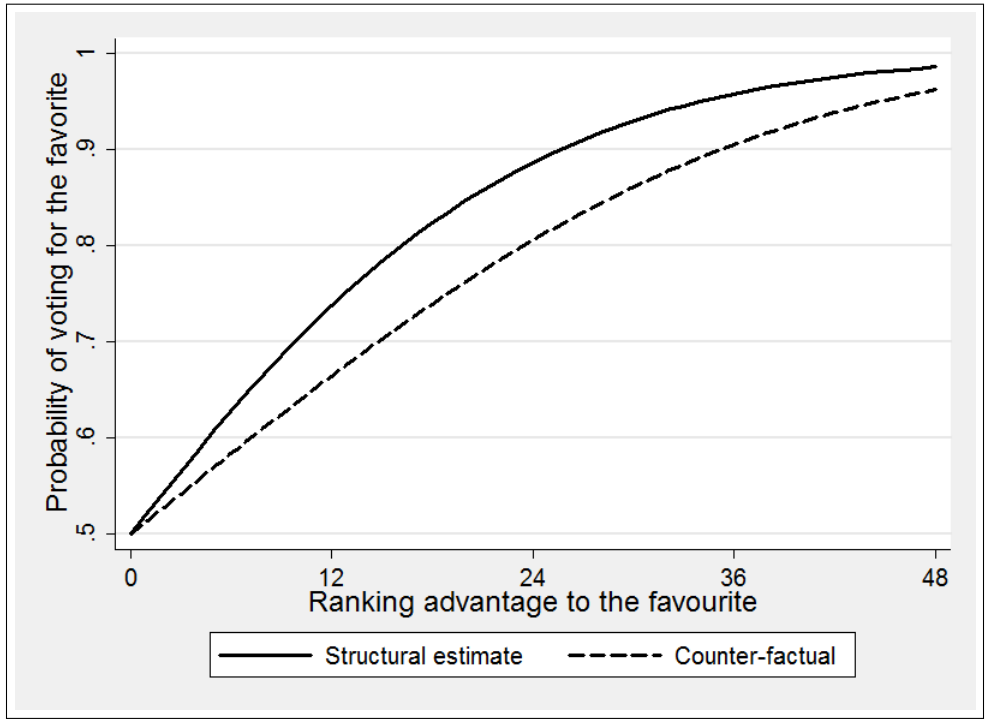


Figure 1.5: ***Probability that the favorite wins: Class 3 judges***

This figure shows the probability that a class 3 judge votes for the favorite, conditional on the class 3 judge having dissented in the previous round; this is the 'structural estimate'. The figure shows how that probability would change if class 3 judges had no preference for coordination (i.e. $\delta_3 = 0$)*; this is the 'counter-factual'.*

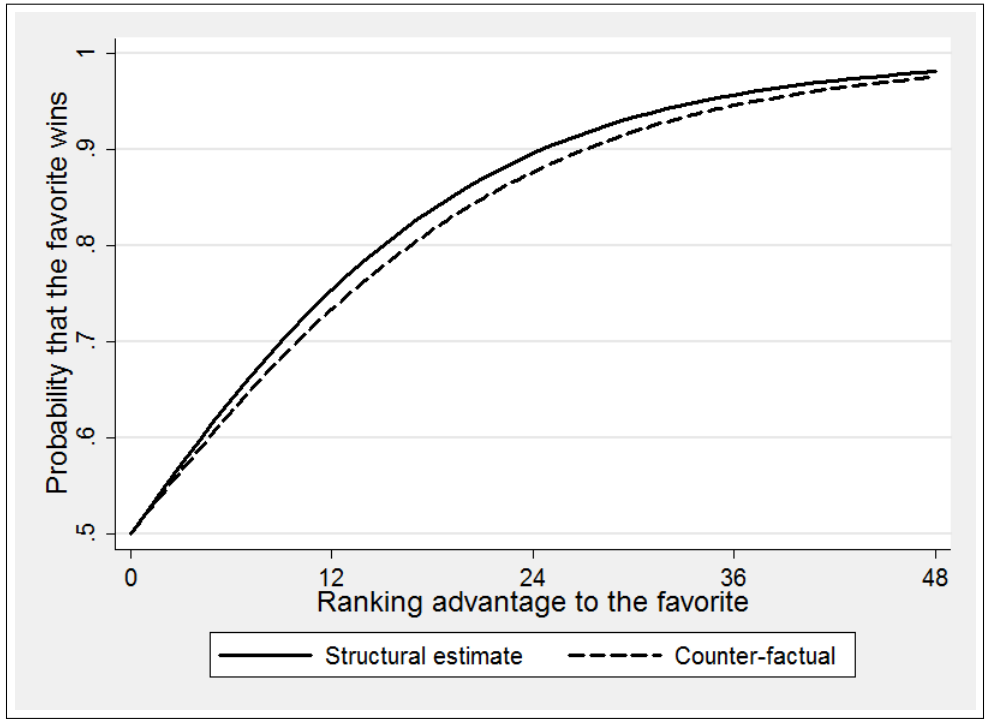


Figure 1.6: ***Probability that the favorite wins: Committee outcome***

This figure shows the probability that the favorite wins, conditional on all three judges having dissented in the previous round; this is the 'structural estimate'. The figure shows how that probability would change if judges had no preference for coordination (i.e. $\delta_1 = \delta_2 = \delta_3 = 0$)*; this is the 'counter-factual'.*

Chapter 2: Global Public Goods, Free-Driving and the Welfare Effects of Geoengineering[1]

2.1 Introduction

The possibility of climatic geoengineering, which is rapidly progressing from fringe science to the mainstream, will produce winners and losers. Current assessments of climate damages suggest that the costs and benefits from global warming would be spread very unevenly across the globe (Intergovernmental Panel on Climate Change (2007); Nordhaus (2010); Tol (2009); Heal (2009)). While moderate amounts of warming may mean higher crop yields and access to faster Arctic shipping lanes for Norway and Russia, the attendant sea-level rise may be enough to wipe out large swaths of densely populated areas in Bangladesh and India. Because climate change will have widely differing effects across regions and countries, the possibility of partially reversing it will also have differing effects, and any realistic assessment of the risks associated with intentionally altering the global climate will have some countries urge restraint at a lower level of action than others.

The difficulty in this situation arises from the non-excludable (equivalently, the non-avoidable) and non-rival nature of geoengineering as an activity. Usually in such public

[1]Co-authored with Jisung Park (Department of Economics, Harvard University)

goods cases the primary problem is free-riding: the under-provision of a public good due to the social benefits of action being greater than the private benefits (a description that dates back to Samuelson (1954)). Geoengineering raises the possibility of a different problem: that of *over*-provision, where if the private cost is low enough and very high levels of action are viewed as harmful by all but a handful of countries, that small set will choose to provide too much of the public good. Given its contrast to the traditional free-rider problem, the most appropriate name for this dynamic would appear to be free-driving.[2]

In this paper, we show that a cheap and unilaterally implementable geoengineering technology may have the potential to transform climate change into a free-driver situation, which, under fairly general conditions, results in a level of geoengineering that is suboptimal from the global social planner's perspective. It is possible that this dynamic will even alter how the world deals with the collective mitigation problem, a topic we return to in the conclusion.

The first contribution of this paper is to develop a formal model of this dynamic. Our model contains the following features: countries differ over the marginal benefit of additional action, each unit of additional action contains some negative side effects for all countries, and there is a private cost associated with taking action. The first two properties combined form what Weitzman (2012) calls a "gob", a public good that can either be "good" or "bad" for an actor, depending on the current level. The third property allows for the possibility of traditional free-riding if high enough, and the possibility of free-driving if low enough. We use this model to motivate a set of characteristics of geoengineering as a physical process that are relevant for their welfare consequences: (1) the cost of geoengineering, (2) the heterogeneity of benefits from geoengineering, and (3) the heterogeneity in assessments of the risks of geoengineering. Our model is intentionally static, yet we also believe that two dynamic properties of the process will be relevant for policy making on this issue: the fact that it is incremental, allowing for the possibility of dynamic games; and it is short-lived and po-

[2]This term is originally due to Weitzman (2012).

tentially reversible. These two additional properties should form the basis of future research.

The second major contribution of this paper comprises a synthesis of the existing literature on the economics of climate change that help us understand these factors. In particular, we summarize different approaches to estimating regional variation in the effects of climate change, which in effect also represent the effects of partially reversing it through geoengineering. We also analyze the nascent literature on the effects of geoengineering, and highlight areas in which further research is needed in order to make informed policy decisions.

Thirdly, we carry out a simple empirical exercise to quantify the likelihood and potential magnitude of efficiency losses resulting from the free-driver dynamic. Using Nordhaus's RICE model (Nordhaus, 2010), we characterize the trade-off between climate change damages and the risks of geoengineering for various regions. The stylized picture that emerges suggests that free-driving is a possibility for a range of plausible outcomes. However, in many of these situations, the damages to countries that would prefer not to geoengineer are greater, at least in financial terms, than the damages from climate change to the countries that would free-drive, and so we suspect that negotiated solutions are readily possible, provided the political architecture is in place for effective negotiation.

We conclude by discussing theory and policy implications that arise from the free-driver dynamic, as well as an urgent future research agenda. One novel theoretical result is that, in contrast to the well-established result that climate damage heterogeneity matters for distributional equity,[3] in the context of planetary geoengineering technologies it matters also for economic efficiency, inasmuch as it determines the gap between private and globally optimal welfare outcomes. Also, unlike in most cases of public good provision, the source

[3]Many think it unfair that those countries hit hardest are those who have contributed least to the problem historically.

of heterogeneity in preferences across countries - that is, whether it is differences in wealth or projected climate damages - actually matters for policy. In the context of climate change and geoengineering, simple cash-transfers can potentially exacerbate the efficiency loss from free-driving, whereas targeted transfers (adaptation support, for example) are more likely to help. Finally, we hope that our simple model of free-driving informs the still nascent search for a global institutional architecture that addresses geoengineering.

The paper is set out in the following way. The remainder of this section acts as an introduction to geoengineering. Section 2.2 sets out a binary version of the free-driving model, with a continuous version reserved for Appendix B.1. Section 2.3 generates best guesses at the relevant heterogeneity in benefits and damages from climate change by synthesizing relevant parts of the existing literature. Section 2.4 then carries out our empirical exercise, attempting to calibrate our model using the best available estimates of the effects of geoengineering and the economic effects of climate change. Section 2.5 discusses policy implications and future research.

An Introduction to Geoengineering

The Royal Society defines geoengineering as "the deliberate large-scale intervention in the Earth's climate system, in order to moderate global warming" (Shepherd, 2009). While many different geoengineering technologies have been proposed (Keith, 2000), we focus our attention on the suite of geoengineering technologies, often referred to as solar radiation management (SRM), which effectively attempt to shade the planet from the sun. Instead of the usual channel of emissions reduction - preventing greenhouse gases from entering the atmosphere - SRM looks to manage the global climate by shooting aerosols into the stratosphere: actively putting new substances in. The simplest way to do this would be to release a relatively small number of sulfate particles into the upper stratosphere, potentially by plane or some sort of projectile cannon (Rasch *et al.*, 2008; Robock *et al.*, 2009). The

reflective properties of these particles would then reduce the total amount of incoming solar radiation, thus offsetting the warming effects of accumulated greenhouse gases in the lower atmosphere.

Planetary geoengineering of this variety is no longer the purview of fringe scientists alone. Formerly excluded from serious discussions of climate science and policy, geoengineering today is quickly entering the scientific mainstream. Some policymakers have begun to consider geoengineering as a hedge against unexpectedly harsh changes in climate (Weitzman, 2009). The US and UK governments have organized task forces to explore the issue in detail, and the topic has been discussed in preparation for the IPCC's Fifth Assessment Report (Edenhofer *et al.*, 2011).

While a comprehensive review of SRM and geoengineering technologies is beyond the purview of this paper, here we highlight three stylized characteristics of SRM that make it likely to feature a free-driver dynamic. First, SRM is potentially very effective in rapidly cooling the earth's climate. Second, the technology is unsophisticated and cheap enough that it can easily be implemented unilaterally. Third, SRM features inherent risks, and a host of known and unknown side effects.

Natural experiments - in the form of large volcanic eruptions like that of Mount Pinatubo in 1991 - suggest that spraying aerosols into the stratosphere can cool the earth quickly and effectively (Robock, 2000). In theory, small amounts of aerosols could achieve large amounts of temperature change. To offset the warming from a doubling of CO_2 concentrations from preindustrial levels, one would have to scatter just two percent of the light that hits the planet (Goodell, 2010; Vaughan and Lenton, 2011). And while there is certainly much room for debate over the relative *precision* of the cooling that could be achieved by SRM, there is very little debate over whether the basic mechanism of planetary cooling by stratospheric particle infusion is scientifically viable.

In principle, planetary-scale SRM could be implemented unilaterally (and cheaply) by a single country, or even a very wealthy private citizen. Due to the rapid rate at which aerosols disperse in the stratosphere, and the relatively small amounts required to achieve substantial cooling, many believe that unsophisticated versions of SRM could be implemented even without the support of an advanced military-industrial complex. Recent engineering estimates put costs using currently available technologies at between 1 and 8 billion dollars per year (McClellan *et al.*, 2012; Robock *et al.*, 2009). Even at the high end, this constitutes less than five percent of GDP for the world's forty largest economies. Any directed technical change would likely lower that figure substantially. Thus, while many may understandably balk at the suggestion that a unilateral actor could tinker with the planet's thermostat, so-called unilateral or "rogue" geoengineering seems at least technically feasible.

Whether or not such unilateral implementation is likely, and what the welfare consequences of such action are for the global community depends in large part on the risks and side effects involved.[4] If SRM did not have any potential side effects, the problem would be much simpler; the only parties that might prefer not to geoengineer would be those who actually stand to gain from anthropogenic climate change. But the reality is that side-effects, including side effects that are potentially catastrophic, are what many believe may make SRM a particularly thorny governance challenge. Known side effects include ocean acidification, massive ozone depletion and its attendant health impacts, changes in regional rainfall patterns and the risk of droughts and floods, alteration of ecosystems due to the effect of dimming on light-sensitive plants, and the fertilization of some plants in a CO2-rich atmosphere (Robock, 2008). Given the planetary scale of intervention, however, it may be the unknown unknowns that are more grave than the predictable effects.

In addition to these risks, SRM technologies feature a dynamic flow property that may

[4]These potential risks are discussed at further length in Section 2.3

complicate matters further. In order to be effective, SRM must be continually applied. Unlike CO_2, a large proportion of sulfate particles will fall out of the sky after a year or so. If they are not injected continually, clearing skies may trigger sudden warming - just like stepping out of the shadows into the sunlight - which may lead to its own adverse consequences (Goes *et al.*, 2011; Baum *et al.*, 2013).

These characteristics - proven effectiveness, unilateral implementability, and the presence of risks and side effects - make planetary geoengineering a potential source of international disagreement. Combined with the underlying heterogeneity in incentives (arising from heterogeneity in potential damages from climate change, as discussed in further detail in Section 2.3) and the as yet complete absence of international governance mechanisms, these features make SRM likely to pose a unique governance challenge. Indeed, engineers, policymakers, and political scientists have hinted at the fact that SRM poses something akin to the converse of the collective action problem arising from emissions-reduction efforts (Victor, 2008; Victor et al., 2009; Virgoe, 2009; House of Commons Science and Technology Committee, 2010; Millard-Ball, 2011; Bodansky, 2011). In the next section, we formalize this intuition, and propose, following Weitzman (2012), to classify it as its own class of public good problems that we call free-driver problems.

2.2 A Simple Model of Free Driving

Weitzman (2012) introduces the idea of public gobs, which are public goods that may be either good or bad, depending on who is consuming it and how much they are consuming. The public nature of the gob is such that everyone consumes the same amount (it is neither excludable nor avoidable), and the gob nature is such that for any given amount, some actors receive positive marginal utility from it, while others receive negative marginal utility.[5]

[5]Conceptually, it is possible to imagine a product that is initially bad for all actors, but then good at sufficiently high levels, and label this product a “bog". This provides the opportunity for amusing wordplay,

Here, we formalize the free-driver intuition in a simple, binary model. Appendix B.1 sketches an analogous continuous model. We postpone discussion of the interaction of this decision with the level of climate change mitigation to the final section.

Consider this problem in the context of N countries facing a binary decision: countries either geoengineer ($g_i = 1$), or they refrain ($g_i = 0$). We might consider the "geoengineer" case as corresponding to a decision to attempt to limit changes in the global temperature to some fixed level, such as the status quo ante of global temperatures at 1992 or 2005,[6] while the "refrain" case corresponds to business as usual.

Each country has a utility function that consists of damages from climate change (D_i) that will be avoided if geoengineering takes place ($G = 1$), a financial cost if they decide to geoengineer themselves (C_i) and a set of possible damages and side-effects if geoengineering takes place, collapsed into a single variable (S_i):[7]

$$U_i = D_i(1 - \mathbb{1}\{G = 1\}) - C_i\mathbb{1}\{g_i = 1\} - S_i\mathbb{1}\{G = 1\} \tag{2.1}$$

For the individual country making a decision in isolation, they will decide to geoengineer if the utility from doing so is greater than the utility from refraining. That is:

$$g_i = 1 \text{ iff } D_i > C_i + S_i \tag{2.2}$$

In isolation, countries for whom the damages from climate change outweigh the financial

but probably yields little additional insight.

[6] An alternative is to assume that the decision is restricting global temperature change to no greater than some amount over pre-industrial temperature levels.

[7] We express these as constants for simplicity, noting that in reality they would reflect a distribution of possible outcomes and costs.

costs of geoengineering, plus the risks associated with the practice, will choose to geoengineer.

Now consider the non-cooperative outcome. Assume that geoengineering has the following production function: geoengineering occurs if any individual country decides to carry it out. Given the rapid and global mixing of sulfate particles in the stratosphere, this seems very likely to be the case (Keith, 2013). That is, $G = \max\{g_i\}$. It follows from (2) that we will have a range of Nash Equilibria in this game, one for each country that would be willing to unilaterally geoengineer. In such an equilibrium, one country geoengineers, and all others refrain to avoid the cost c_i. If we allowed for cost-sharing, we can imagine other equilibria in which coalitions gather to geoengineer, subject to free-riding concerns.

Next, let us compare this to the socially optimal solution. Assuming all countries have equal weight in the social welfare function, we have:

$$U = \sum^{N} D_i(1 - \mathbb{1}\{G = 1\}) - C_i\mathbb{1}\{g_i = 1\} - \sum^{N} S_i\mathbb{1}\{G = 1\} \tag{2.3}$$

The social planner would then choose to geoengineer under the following circumstances:

$$G^* = 1 \text{ iff } \sum^{N} \frac{D_i}{N} > \frac{min\{C_i\}}{N} + \sum^{N} \frac{S_i}{N} \tag{2.4}$$

That is, geoengineering is optimal if the average damages from climate change are sufficiently large to merit taking on the financial costs of SRM in addition to the average expected risk of adverse side effects. Comparing (2) and (4) allows us to elucidate the ideas of "free-riding" and "free-driving" in a binary context.

Free-riding occurs if it would be socially optimal to geoengineer but no individual

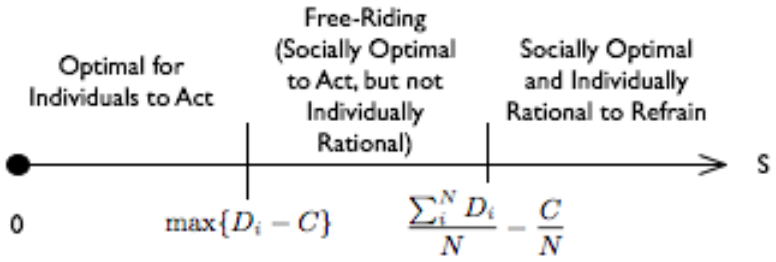


Figure 2.1: ***Free Riding with homogenous C and S***

country is willing to bear the cost to carry it out: $G = 0$, $G^* = 1$. This will occur if the cost of geoengineering (C_i) is sufficiently large relative to the benefits (in terms of avoided damages from climate change, D_i), controlling for possible adverse side effects (S_i):

$$\sum^N \frac{D_i}{N} > \frac{C_i}{N} + \sum^N \frac{S_i}{N} \text{ but } D_i > C_i + S_i \forall i \tag{2.5}$$

If we consider side-effects to be global and common in nature ($S_i = S \; \forall \; i$) and that all countries have access to the same technology ($C_i = C \; \forall \; i$), then this condition collapses to $\sum^N \frac{D_i}{N} - \frac{C_i}{N} > S > D_i - C$, and can be depicted graphically by varying S (Figure 2.1).

In contrast to the free-rider case, there are situations in which it would be optimal to refrain from geoengineering, but individual countries will nonetheless find it in their private interest to go forward. We call these situations “free-driving", in the sense that the country that feels most strongly about G drives the global level: $G = 1$, $G^* = 0$.[8] This will occur if C is sufficiently low, and there is sufficient heterogeneity in D_i and S_i.

$$D_i > C_i + S_i \text{ for some } i \text{ but} \sum^N \frac{D_i}{N} < \frac{C_i}{N} + \sum^N \frac{S_i}{N} \tag{2.6}$$

[8]Note that, in the continuous case, the free-drivers are the set of nations who contribute to the level of G. See Appendix B.1.

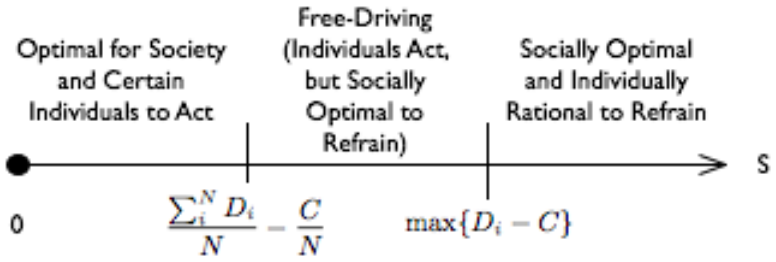


Figure 2.2: ***Free Driving with homogenous C and S***

Note that if $D_i > S_i$ for all countries, then free-driving is not possible - in such a world geoengineering would be a public good, with no "bad" aspect of the public good involved. If we again assume a common S and C, the condition collapses to $D_i - C > S > \sum^N \frac{D_i}{N} - \frac{C_i}{N}$, which we can again depict diagrammatically (Figure 2.2).

For the free-driving case, if we assume that all countries have access to the same technology and that its cost is vanishingly small ($C_i = 0 \; \forall \; i$, as in Weitzman (2012)), then all that is required for geoengineering to occur is that *any* country believes that a geoengineered world is superior to a world with climate change. It is this sense in which a very cheap public technology, combined with heterogeneous preferences, is highly likely to result in free-driving. This is the reverse of the free-rider case: there, heterogeneity makes it more likely that one nation will perceive it as sufficiently in their private interests to provide a good that everyone collectively would prefer provided, while here it makes it more likely that one nation will unilaterally provide a good where the collective would prefer to see restraint.[9]

While this is obviously a very simple binary game, it makes two points. The first is the possibility of a free-driving effect, with potentially large welfare implications. Secondly, it points us in the direction of the parameters we need to know in order to assess the changes

[9]The continuous case in Appendix B.1 makes this more starkly.

of free-driving taking place: the cost of geoengineering, the heterogeneity in the damages from climate change, and heterogeneity in the assessment of the risks of geoengineering. In the next section, we turn to estimation of those parameters.

2.3 Estimating Differences in the Effects of Climate Change and Geoengineering

The overarching theoretical framework in which climate damage estimates are situated is seemingly simple: compare the costs of mitigating greenhouse gas emissions to the benefits in terms of future damages avoided. As many have pointed out however, cost-benefit analysis in the context of climate change is anything but simple. Three aspects of climate policy greatly complicate the analysis, making climate change what Martin Weitzman calls "the problem from hell": long time-horizons and inherent lags which necessitate inter-generational discounting; the global scope of the problem and related questions of intra-generational equity; and pervasive uncertainty, including both fat-tailed distributions and significant ambiguity. Each of these problems has generated a significant literature of its own.

Because of the pure global public good nature of climate mitigation, academic economists and policymakers have been mostly concerned with the aggregate social cost of carbon: that is, the weighted mean over all countries and individuals. As a result, relatively little attention has been paid to clarifying the precise nature of the heterogeneity in climate-related damages for different countries and regions. However, as the model in Section 2.2 makes clear, in free-driving contexts it is the right-hand tail of the distribution of damages across countries that matters most. To the extent that regional variation has been investigated, it has typically been from the perspective of equity, rather than efficiency.

In this section, we provide an overview of the current state of the literature on this narrow slice of climate change economics. We do not attempt a comprehensive literature review here.[10] Rather, we propose a new lens through which one might view the damage estimates to date: one which stresses the dimension of damage heterogeneity as opposed to average or aggregate damages.

2.3.1 Methodology

The means of estimating the regional effects of climate change mirror the approaches the literature takes in measuring the economic impacts of climate change more generally. Typically, any estimation of this sort requires three things: a model or a set of assumptions about future emissions, a model of the relationship between emissions and the global climate (typically including temperature, sea level rise and changes in precipitation), and a means of mapping those climatic changes to economic outcomes. While there is substantial uncertainty and variation in approaches to each requirement, our focus is on the third step. Most studies about that question follow one of two major approaches: the bottom-up or "enumerative" approach; and the top-down or "statistical" approach.

The bottom-up approach examines the potential impact of a changing climate on specific sectors or causal channels (e.g. agriculture, sea-level rise) and aggregates these effects one by one. The examined sectors primarily include agriculture (Mendelsohn *et al.* (2000); Schlenker and Roberts (2008)), health (Deschênes and Greenstone (2011)), and impacts of sea-level rise on major infrastructure (Stern (2007)), although as research has progressed greater numbers of sectors have been included. These sector- and region-specific estimates are extrapolated to other regions, then aggregated over time and space, based on various projections for future climate change taken from climate models.

[10] A useful summary is Tol (2009).

The top-down approach examines the observed macroeconomic relationship between climate variables and income directly, and extrapolates overall damages based on scientific models of future climate change (see for example Rabl and van der Zwaan (2009)). Studies such as Nordhaus (2006), which utilize geographically referenced data on a grid-cell basis, also fall under the latter heading but at a lower level of aggregation. This approach involves more abstraction from specific mechanisms, but is driven by observed macro-statistical relationships between climate and income aggregates. For example, Dell *et al.* (2009) find a correlation between hotter-than-average years and lower-than-average GDP growth on a country-level panel: approximately -1.1% GDP per +1° C of warming. But they are agnostic as to the causal mechanism, and do not explain why the relationship holds only for a subset of "poor" countries. Indeed, a key weakness of this approach is that few of these studies posit a direct causal relationship between temperature and economic output or welfare (Heal and Park, 2013). Moreover, using observed climate-economy relationships to impute potential future damages requires somewhat unrealistic assumptions about adaptation (e.g. changes in prices and production methods in response to future climate change).

Each of these approaches has strengths and weaknesses. The bottom-up approach is based on specific proposed channels, and so typically uses studies exploiting quasi-experimental variation and formal models of processes. However, this approach is also subject to criticisms that extrapolating measured damages in one context to another can involve substantial errors (Brouwer and Spaninks (1999)), that it may omit entire mechanisms for damage propagation, and that it can easily underestimate adaptation. By contrast, the top-down approach is based on cross-country analysis from the start, and so avoids some of the difficulties of out of regional sample extrapolation. It is, however, subject to the concerns that it often does not provide any sense of what causal mechanisms are at work, and may face difficulties in considering factors which have little historical variation, but may have great future variation, such as sea level rise and carbon dioxide fertilization. The best approach, therefore, would appear to be using a blend of these two methods.

2.3.2 Variation in Climate Change Impacts

Table 2.1 summarizes models that generate estimates of the total effects of climate change by region. We can see that the literature predicts great regional variation in climate changes damages, and also significant differences in which regions will be hardest hit across models. At mild levels of temperature change, such as an 1°C, the literature suggests that most regions will gain, with serious negative consequences restricted to Africa and, to a lesser extent, South Asia. At more significant levels of climate change, such as 2.5°C, most countries are predicted to suffer from net negative consequences, with net benefits restricted to the high latitude countries of North America and Russia. The largest negative consequences are again predicted to be felt in Africa and South Asia. This general pattern of relative effects has been recognized by the IPCC and Stern Reviews. Note that none of these particular studies directly consider the potentially catastrophic effects of greater levels of climate, a topic upon which the literature has struggled to generate rigorous estimates.[11]

What drives these patterns in variation in climate change damages we observe? Three factors seem most closely linked to high levels of vulnerability. The first is that already heat-stressed regions are likely to be hit hardest by future warming, as any given temperature increase will likely have disproportionately damaging effects, as societies are pushed closer to the limits of human habitation. A factor pushing in the other direction is that countries in lower latitudes will likely see smaller increases in temperature (see, for example, Manne *et al.* (1995)). However, it is generally considered that this effect is not large enough to outweigh the negative effects of an already high temperature.

The second factor connected with high levels of damage is vulnerability to the most significant effects of climate, such as agriculture, sea level rises, water availability and

[11]On the theoretical modeling of catastrophic effects, see Weitzman (2009).

disease. A good demonstration of this is contained in Bosello *et al.* (2012), who consider regional variation in climate changes damages by type of impact. Bosello *et al.* find that Africa, East Asia (excluding China), India and South Asia suffer the greatest losses from climate change, and that this is primarily driven by changes in agriculture. On the other hand, the Former Soviet Union, China and Northern Europe benefit from moderate climate change, this time from beneficial changes in agriculture. India and South Asia also suffer disproportionately from sea level rise.

The third factor is low levels of economic development. Low levels of economic development are correlated with the two previous factors, but also suggest weak institutions, and therefore difficulties with adaptation. This point is made frequently in the literature (see Adger (2006); Alberini *et al.* (2006); Smit and Wandel (2006); Tol (2008); Tol and Yohe (2007); Yohe and Tol (2002))), and so gives yet another reason to believe that strong institutions are crucial for economic growth in the coming century. An acute famine in Northern Europe may, for example, lead to temporarily elevated prices and a deteriorated trade balance; the same famine in a South Asian country with weaker institutions may lead to riots and violent conflict.

The combination of these three factors leads to the conclusion that it will be the developing world that will suffer the most from climate change, significantly more than in the developed world. At a national level, this suggests that the most affected countries will be island communities such as Pacific islands and the Maldives, or very low lying developing countries like Bangladesh. At a regional level, the regions that are likely to suffer the greatest damages from climate change are South Asia and Africa, much more so than China or Latin America.

There is a dissenting argument to this point of view. This argument suggests that willingness to pay for environmental amenities, and also the statistical value of a human

life, are likely to be increasing in income. As a result, the absolute losses may be greatest in the developed world (see, for example, Manne *et al.* (1995)). However, it is likely that that is not the best measure of the heterogeneity this paper is interested in - we primarily want to focus either on losses as a fraction of income, or to map losses into a welfare- or utility-based metric.[12] If we do that, we will find ourselves back with the conclusion that it will likely be the developing world most interested in geoengineering technology from a free-driving perspective.

It is also worth noting that the literature studied here, which consists of aggregated studies at the regional level, likely reflects a rather conservative view of the degree of damage heterogeneity. If we instead consider studies that look at country-specific impacts for particular sectors or climate processes, we often observe much larger levels of heterogeneity. For example, Hsiang and Narita (2012) estimate that the value of lost GDP due to increased storm frequency and intensity varies enormously across affected countries. Schlenker and Roberts (2008) show that the agricultural damages from climate change also vary enormously across countries, with some areas in northern latitudes benefiting from longer growing seasons and tropical and semi-arid areas suffering diminished crop yields of over 75% by 2050. The fact that most damage functions take some measure of annual average temperature as the independent variable also suggests existing estimates are likely to be conservative. As emerging work using more detailed measures of temperature (for example, the number of cooling degree days and heating degree days) has shown, much of the action in terms of welfare impacts from climate-related events comes from a few days of extreme heat (or cold) during the course of the year (Deschênes and Greenstone, 2011; Schlenker and Roberts, 2008).

[12]Noting, of course, the risks of associated with interregional comparisons, and with the utilitarian framework more generally. See Sen and Williams (1982).

2.3.3 Estimating the Effects of Geoengineering

In our analysis of geoengineering as a free-driving problem, the twin consideration to heterogeneity in climate change damages is heterogeneity in the effects of geoengineering. Unfortunately, at the moment we know relatively little about the potential effects of planetary geoengineering, let alone heterogeneity in those effects. There are, however, several expected side effects, with some significant regional variation.[13]

Known side effects of SRM include ozone depletion, as other particles are injected into the stratosphere (Crutzen, 2006; Tilmes *et al.*, 2008); alteration of ecosystems for a range of reasons, including the impact of dimming on light-sensitive plants, the availability of water, and fertilization of some plants in CO2-rich atmosphere (Mohan *et al.*, 2006; D'Arrigo *et al.*, 2008); and possibly acute changes in rainfall, with attendant risk for drought and floods (Liepert *et al.*, 2004; Oman *et al.*, 2006; Trenberth and Dai, 2007). Each of these likely has a regional component, in particular potential changes in precipitation (Matthews and Caldeira, 2007). Robock *et al.* (2008) suggests that planetary level SRM may be sufficient to disrupt the Indian Monsoon, with potentially devastating effects for South Asia. It is also possible that SRM will alter the probability of el Niño events (Adams *et al.*, 2003).

Another dimension of risk associated with geoengineering is the dangers associated with halting geoengineering once it has been in use. If SRM is used to limit temperature increases from global warming, then the short-lived nature of the technology implies that if the injection of aerosols is reversed, temperatures will rapidly increase (Brovkin *et al.*, 2009; Ross and Matthews, 2009; Matthews and Caldeira, 2007). Baum *et al.* (2013) raises the possibility of a "double catastrophe", in which some societal-level threat such as a global war or pandemic also leads to cessation of geoengineering, compounding the initial precipitating event with rapid climate change. In this sense, geoengineering potentially

[13] A good non-technical summary of the potential issues is given by Robock (2008).

raises the stakes of other global issues, a risk that some nations may take more seriously than others.

Each of these is in addition to the known effects of high greenhouse gas concentrations in the atmosphere that SRM does not have the potential to ameliorate. In particular, even if SRM achieves "optimal" temperatures successfully, if greenhouse gas emissions are not reduced then ocean acidification will still occur, with significant impacts on the world's tropical coral reefs (Shepherd, 2009; Feely *et al.*, 2004)). This suggests, amongst other things, that SRM is not a substitute for mitigation, particularly without additional means of addressing greenhouse gas concentrations.

How do these potential side effects map into aggregate damages and human welfare? Unfortunately, at this stage we do not have studies that provide aggregate estimates of the sort that exist for climate changes, both because this is a newer field of study and because the effects are dramatically more uncertain. This is a serious problem for our ability to make predictions about the free-driver dynamic in this area, and leads us to take a particularly simple approach to side effects in the next section. It also suggests that this should be an area of great research urgency, both to help inform policy-makers about the best courses of action with respect to geoengineering and to help us better understand the possibility for a free-driver dynamic.

2.3.4 From Climate Damage Heterogeneity to Free Driving Scenarios

The literature summarized here suggests that there will be significant variation in climate change damages. From the current evidence it appears that Africa and South Asia are the regions most likely to want to avail themselves of any technology that might reverse or forestall climate change. The nations in these regions have the combination of already hot temperatures, high likelihood of the most extreme impacts of climate change and weak

institutions that makes them highly vulnerable to the changes associated with a warming world.

On the other hand, not enough is known about the possible effects of geoengineering to give an informed statement about which countries and regions will be most (and least) concerned about pursuing the technology. It appears that there will be regional variation in the effects of any geoengineering technology, but we do not yet have precise estimates of those effects. In general it seems that if a free-driver is to emerge, it will be from Africa, South Asia or perhaps one of the low-lying island nations. However, if we take concerns about the effects of geoengineering on the Indian Monsoon cycle seriously, that may convert India from a potential free-driver to one of the largest losers from any such dynamic.

The next set of questions we want to ask is under what circumstances is a free-driver dynamic likely to evolve, and what the welfare consequences of such a dynamic would be. For that, we need a finer model of the relationship between climate changes and welfare effects to fully characterize the trade-offs. In the next section we use a specific integrated assessment model to provide an illustrative first-pass at addressing these questions.

Before we move on, however, it is worth emphasizing again how incomplete our state of knowledge is in this area. There are a number of dimensions along which our analysis of the potential impacts of climate change and geoengineering may be dramatically incorrect. These include the possibility of missing channels for climate to affect welfare altogether, the fact that many of the models discussed above do not explicitly include any element of risk or stochasticity,[14] and the absence of consideration of catastrophic tail-events in most of these models.[15] The literature summarized here is also typically comparative static in

[14]One exception is the PAGE model (see Hope (2006)). Stern (2007) cites this as a reason for his use of the PAGE model.

[15]See Weitzman (2009).

nature: as Tol (2009) points out, these estimates are generated by imposing future climate on today's economy and by considering adaptation using simple assumptions, if at all. The addition of true dynamics makes estimates in this area even more uncertain.

Finally, we have assumed that the appropriate actor in this area is the nation state, and so neglected variation within countries. Research on intra-country variation on effects is rare, particularly outside of the United States, with O'Brien *et al.* (2004) being one of the few exceptions. However, it is likely that climate change effects would not be homogeneous within countries; certainly, particular economic sectors (such as agriculture), regions (coastal zones), and age groups (the elderly) are more heavily affected than others. This has implications for any political analysis of climate related policy.

2.4 Empirical Exercise

2.4.1 The RICE model

The RICE (Regional Integrated Climate and Economy) model is an extension of Nordhaus's DICE (Dynamic Integrated Climate and Economy) model that includes different regional effects and behaviors.[16] We use the RICE2010 model, available at Nordhaus's website.[17] RICE and DICE both view climate change from a classical growth theory perspective, with different regions investing in capital and climate investments (abatement), gaining higher consumption in the future in exchange for lower consumption in the present. Capital affects future consumption through a Cobb-Douglas production function, while emissions form a "negative natural capital" which lowers future production through a geo-physical model and estimates of the effects of global temperature on production.

[16] A good summary of the RICE model which is very closely related to the discussion here is Nordhaus (2010).

[17] `http://nordhaus.econ.yale.edu/RICEmodels.htm`

The RICE model divides the world into 12 regions. Some are large countries such as the United States and China, while others are regions consisting of many countries such as the European Union or Africa. Each region is assumed to optimize a well-defined social welfare function by selecting consumption, emissions levels and investment at each period of time. The social welfare function features diminishing marginal utility of consumption, in practice taking power utility form. Each region discounts future periods using a pure rate of time preference, which is treated as equal across all regions. The curvature of the utility function and the rate of time preference are calibrated so that the real interest rate in the model is close to the observed average real interest rate and average real return on capital.

The economic part of the model consists of these 12 regional economies, each producing a single commodity that can be used for consumption, investment or emissions reduction. Each region is endowed with an initial stock of capital and labor and an initial level of technology. Population growth and technological change are exogenous, with population figures and projections taken from the United Nations.[18] Capital accumulation is endogenously determined by optimizing the flow of consumption over time. The model calibrates the parameters using data on historical GDP and CO_2 emissions for 1960-2008, and the emissions reduction cost functions are drawn from more detailed models in the IPCC Fourth Assessment Report and the Energy Modeling Forum 22 Report. The model also assumes the existence of a relatively expensive backstop technology that can replace all carbon fuels at a sufficiently high price.

The geophysical part of the model consists of a number of equations that describe a simplified version of the relationship between various factors that affect climate change. These include emissions, the carbon cycle (using a three-reservoir model), radiative forcing, a climate model, sea level rise, and regional climate-damage relationships. The model

[18]United Nations (2004).

contains both endogenous emissions, being industrial CO_2, and exogenous emissions from land-use changes, non-CO_2 greenhouse gases and sulfate aerosols. This element of the model is based on simplifications of more complicated models.[19]

The key part of the model for our purposes are the regional climate-damage relationships. In the RICE model, a fraction of GDP $\Omega^k(t)$ in country k and period t is lost to climate change damages. This $\Omega^k(t)$ is a function of the global mean surface temperature $T(t)$ (and, in some versions, sea level rise $SLR(t)$), and is essentially expressed as the sum of a quadratic function and a catastrophic damages function in the following form:

$$\Omega^k(t) = \frac{g^k[T(t), SLR(t)]}{1 + g^k[T(t), SLR(t)]^\omega}$$
$$g^k[T(t), SLR(t)] = \psi_1^k T(t) + \psi_2^k T(t)^2 + \psi_3^k (T(t)/T_{cat})^{\psi_4}$$

T_{cat} is the threshold temperature for catastrophic damages. In the baseline runs of the RICE model, and in all runs used in this paper, the possibility of catastrophic damages is turned off, with ψ_3 set to 0.

Where do these damage functions come from?[20] They are constructed by using the enumerative, bottom-up approach described in Section 2.3. Nordhaus uses damage estimates from the literature for range of categories: agriculture, sea level rise, health, non-market amenity impacts, human settlement and ecosystems and catastrophic damages. For each of these categories, two scenarios are considered: a 2.5°C warming scenario, and a 6°C warming scenario. The total damages for each region are then aggregated for each scenario.

To generate the specific functions used here, Nordhaus takes a quadratic approximation

[19] See Nordhaus (2010) for more details.

[20] This process is described in Nordhaus and Boyer (2000) for the RICE-99 model, and the current RICE model uses a similar approach.

using the three points of the status quo, the 2.5°C scenario and the 6°C scenario. The functional form described above is then used to bound damages above by 100%. The quadratic form of damages used in the RICE model is, of course, a dramatic simplification, but gives very stark and easily interpreted results for our purposes. As discussed in Section 2.3, models of this sort often generate conservative estimates of damages, particularly for very large temperature changes.

One downside of this approach is that by assumption all countries suffer damages for all levels of climate change, whereas the best estimates as discussed in Section 2.3 suggest that some countries stand to gain from mild warming. For our purposes, this means that, in the absence of side effects or other damages, all countries will agree to geoengineer, at least until the climate is returned to the pre-industrial temperature levels. We tackle this problem by first analyzing the heterogeneity in damages, then introducing a simple way of thinking about side effects, such that some countries with low levels of damages would prefer to avoid potential side effects and therefore disagree over geoengineering with those countries that are willing to suffer the side effects to ameliorate large climate damages.

As is noted in Nordhaus (2010) and elsewhere, solving a multi-country general equilibrium model is difficult for a number of reasons, one of which involves normative assumptions implicit in the social welfare functions used. Nordhaus uses a modification of the Negishi procedure introduced by Nordhaus and Yang (1996). As is typical in the Negishi procedure, optimization involves weighting the utilities of the different regions to equate marginal utilities in the initial period, in effect accepting the present distribution of wealth as a socially optimal baseline. This enables global optimization without generating large intraperiod transfers of wealth from wealthy regions to poor regions. However, we do *not* use these weights, because a) our interest is positive, rather than normative, and we are not seeking to optimize global behavior, and b) to re-weight utilities in this fashion would be to impose an answer to the question we are asking.

2.4.2 Heterogeneity in Climate Damages

Table 2.2 sets out the parameters for the 12 regions in RICE 2010. At this level, one can notice some patterns in the predicted damage functions. As a fraction of GDP, Africa, India and other non-OECD Asia have the most to lose from climate change, with the highest linear and quadratic coefficients. China and Latin America have linear components to damage, but not particularly high quadratic coefficients. Russia faces the lowest damages, while other countries that we suspect will have low damages are grouped together into either EU or Other High Income countries, and so are not immediately apparent. Note that these parameters omit catastrophic effects, and hide a lot of variation within regions (compare, for example, the likely effects of climate change on Australia and Canada within OHI, or Nepal and Bangladesh within Other non-OECD Asia).

The parameters in Table 2.2 do not tell us much in and of themselves: we want to convert those damage multipliers into damages in dollar terms, and then into utility terms, noting that the same dollar loss is of much greater significance to poorer regions. Using 2005 GDP figures and the power utility function used in the RICE model (with a curvature parameter of 1.5), Table 2.3 and Figures 2.3 and 2.4 express damages from a range of possible temperature changes.

In terms of GDP, it is the wealthiest and largest regions of the US and the EU that suffer the most from moderate to high levels of climate change, despite having lower multipliers. It is perhaps unsurprising that the dramatically larger GDPs of those regions outweigh, in dollar terms, the larger proportion of damages to the (as of 2005) smaller economies of China, India and Africa. This suggests that, in terms of dollars to be gained, the US and EU will be the most in favor of geoengineering. Note also that we have yet to consider risks of side effects or geoengineering-induced climate catastrophes.

However, the higher per capita GDP of the US and EU mean that those regions also have

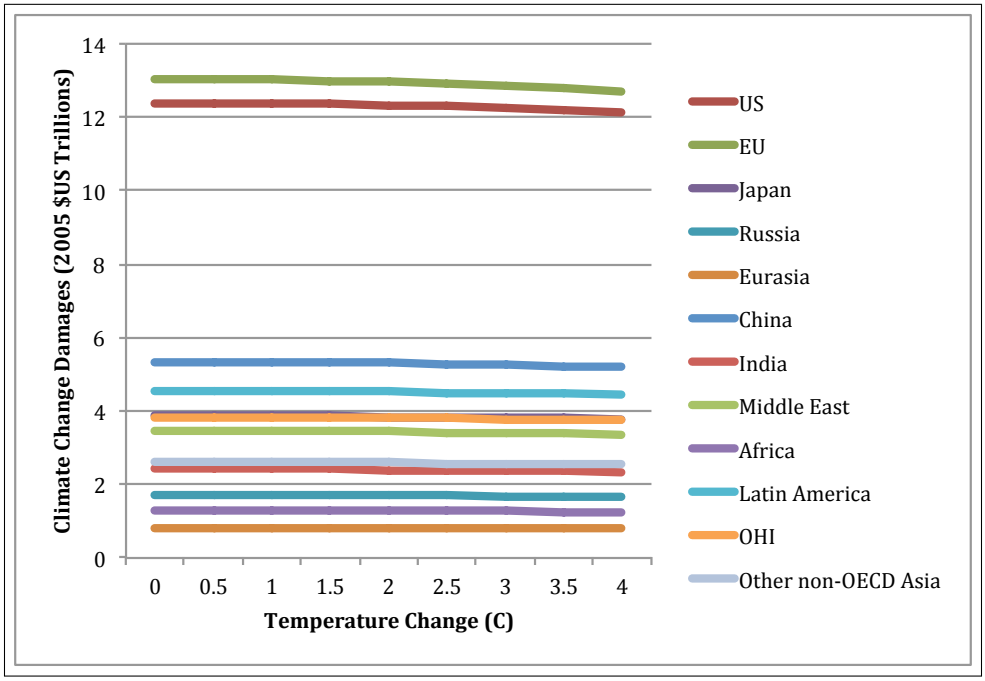


Figure 2.3: *GDP at various levels of temperature change*

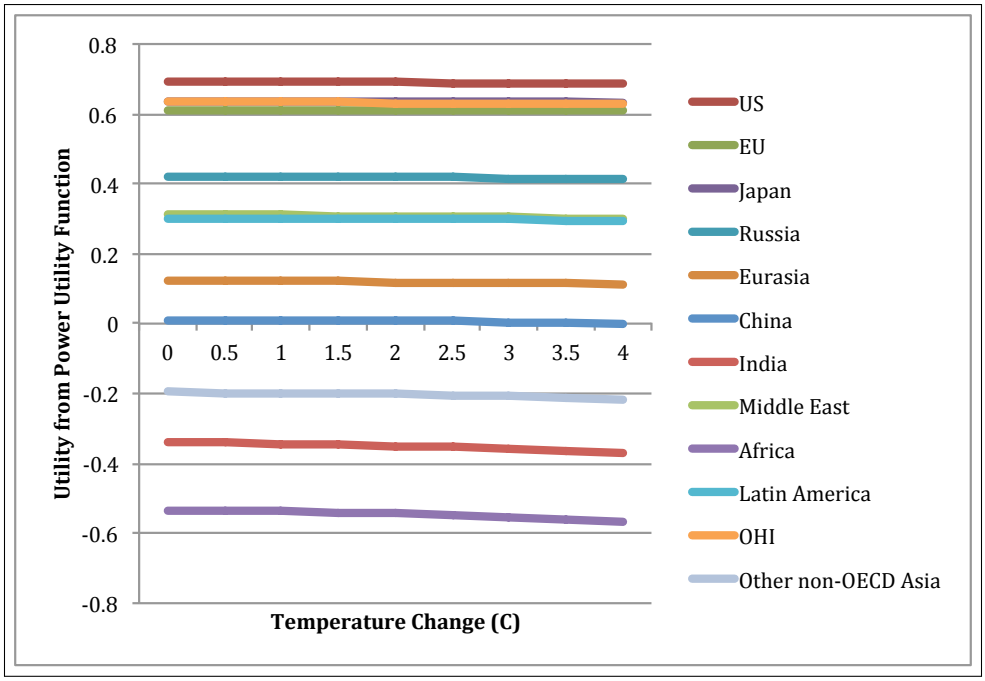


Figure 2.4: *Individual utility at various levels of temperature change*

very low marginal utilities of consumption, and as Figure 2.4 shows, the smaller GDP losses of the poorer regions lead to larger utility losses in those areas. Africa, India and the other non-OECD Asian nations suffer greatly from high levels of temperature increases, while the US, EU and the Other High Income countries do not have large changes in utility at all. From this perspective, it is the poorer regions of the world that will be most in favor of geoengineering.

The results in Table 2.3 and Figures 2.3 and 2.4 indicate a wide variation in potential gains from geoengineering, yet are incomplete because they are simply snapshots based on current economic conditions. A fuller answer would look at damages and gains over the foreseeable future, taking into account economic and population growth and the likely timeline of temperature changes. To calculate those effects, we need to use the full RICE model, which we turn to next.

2.4.3 Using the full RICE Model

Before discussing the results from the full RICE model, it is worth quickly discussing how we carry out our analysis. The RICE model has regions optimizing consumption and abatement decisions, then iterates those until convergence on a Nash Equilibrium. Our results here rely on countries carrying out that optimization given the effects of climate change under the business as usual (BAU) scenario, and then comparing output, consumption and utility with and without climate change, assuming that geoengineering can completely and perfectly undo the effects of climate change, while holding investment decisions constant. It would be possible to have regions re-optimize their saving decisions in the presence of geoengineering, with regions saving more as geoengineering increases the future marginal product of capital. However, this will not enable us to compare gains to geoengineering as cleanly, and will likely not significantly affect the results.

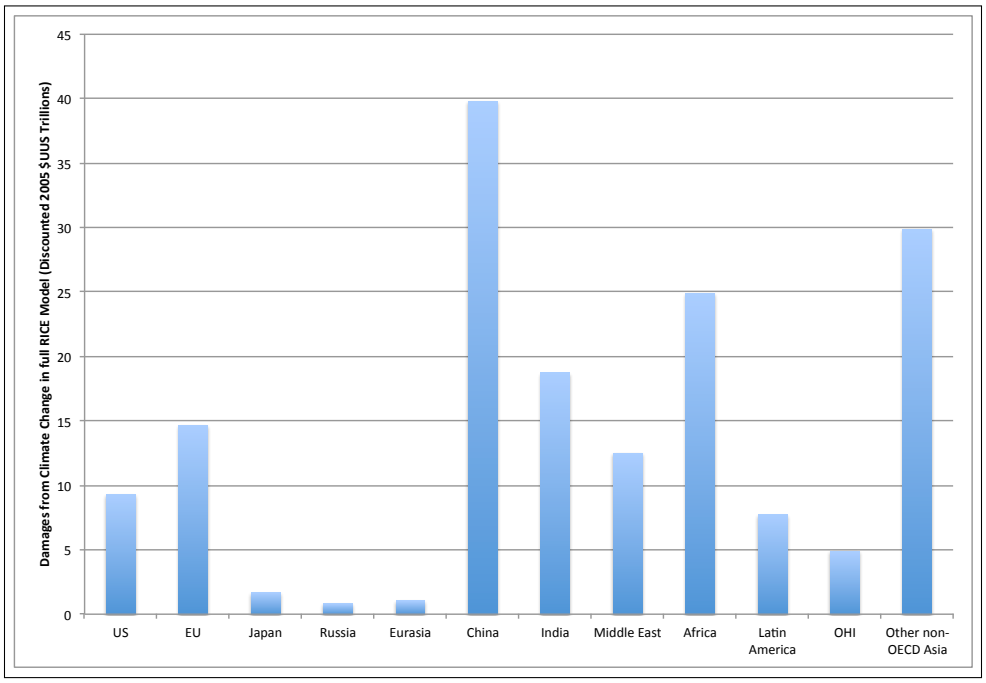


Figure 2.5: ***Difference in GDP between BAU with and without climate damages***

In GDP terms (Figure 2.5), note that in contrast to the results based on 2005 GDP, it is now China that loses the most GDP over the full time horizon of the RICE model. This difference comes from the fact that the model has China grow so much over the next 300 years. The next largest losers are the regions with the largest damage multipliers: Africa, other non-OECD Asia and India. In contrast, Russia, Japan and Eurasia have very limited losses over the full horizon, and so would appear to have relatively little interest in geoengineering.

In terms of per capita utility (Figure 2.6), Africa suffers the most, through a combination of relative poverty and large monetary losses. Next come India and China, which while wealthier than Africa, are also relatively poor and have relatively large monetary losses compared to the rest of the world. If we aggregate per capita utility into national utility (Figure 2.7), which is effectively to weight by population, then the order of losses stays the same, but the magnitude of losses changes, such that the large losses to Africa, China and India completely dwarf the relatively small losses to the wealthy world.

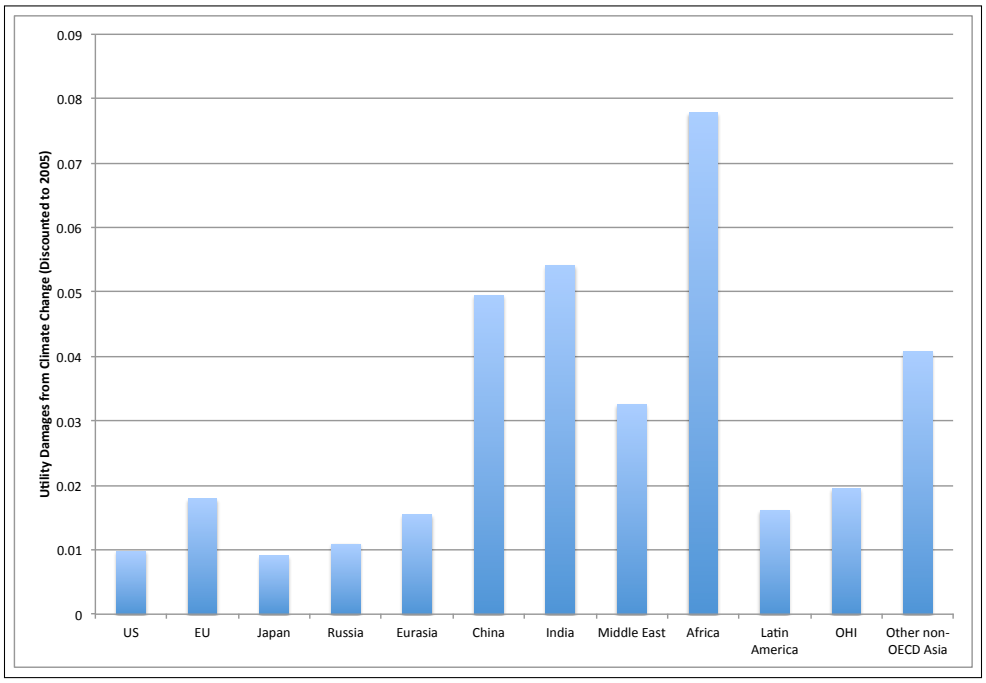


Figure 2.6: ***Difference in Utility between BAU with and without climate damages***

At this stage, we have demonstrated heterogeneity in the benefits from geoengineering (potentially avoided costs from climate changes). However, because all countries have quadratic damage functions in temperature, without further further, they all prefer geoengineering to not. To generate a possible difference of opinion, we introduce S_i, the potential side effects of attempting to geoengineer.

2.4.4 The Risks of Geoengineering and the Potential for Disagreement

There are two costs associated with geoengineering: financial costs and risk of side effects, both known and unknown. As discussed in Section 2.1, the financial costs of many forms of geoengineering, in particular SRM, is expected to be well within the reach of individual states and even wealthy individuals. To simplify analysis, and to avoid the possibility of free-riding effects if coalitions must form to fund geoengineering, for the purposes of this section we assume that it is effectively free.

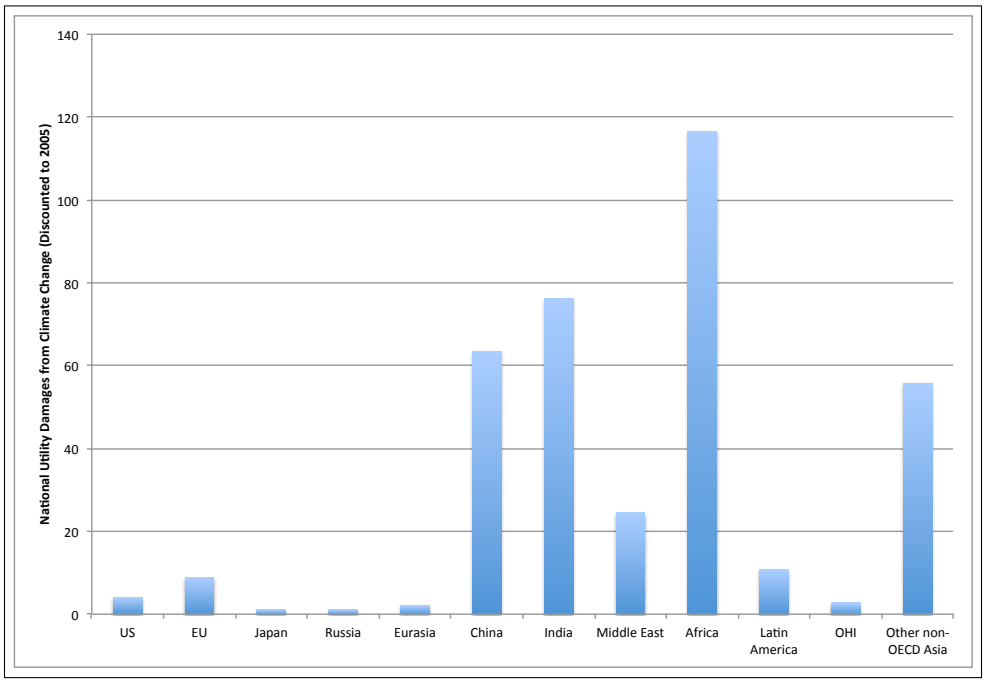


Figure 2.7: ***Difference in Utility aggregated at the Regional level between BAU with and without climate damages***

This leaves us with side effects and other damages as the key source of disagreement over the decision to geoengineer. As mentioned in Section 2.3, the potential damages include ozone depletion, disruption of ecosystems and changes in precipitation. These side effects are currently not well understood, either in their nature or their potential effects on output or human happiness at a global, regional or state level. Rather than attempt to get ahead of the science, and suggest damage functions and probability distributions for these side effects, we will simply assume that states are capable of transforming the best science about those risks into a certainty equivalent cost associated with geoengineering. This is a dramatic simplification, but it enables us to generate stark results. Much more research is needed on this question, from both the physical sciences and the social sciences.

More concretely, we assume that the risks of geoengineering represent a fraction α of GDP

for each region.[21] Each region k therefore chooses between two options: geoengineering, with climate change prevented but regional output reduced in expectation to $(1-\alpha)GDP_k$, and refraining, in which case they avoid the risks associated with geoengineering but incur the costs of climate change discussed above. We then vary the geoengineering risk factor α, and compare it with various levels of temperature change and associated damages from the RICE model to examine the possibility for disagreement and free-riding.

The trade-offs are presented in Table 2.4, using 2005 GDP as the relevant measure of output.[22] Each cell corresponds to a value of α (x-axis) and temperature change (y-axis) pair for each country, and is the difference between GDP under geoengineering and GDP under climate change. Positive numbers, marked in green, suggest the region would prefer to geoengineer, while negative numbers, marked in red, suggest the region would prefer restraint.

Table 2.4 points to the possibility of free-riding dynamics. Note that at high temperature changes and low costs of geoengineering (the bottom left for each country) all countries prefer to geoengineer, while at low temperature changes and high costs of geoengineering (the top right for each country) all countries prefer restraint. However, the central squares for each country are where disagreement lies - where the costs of climate change and geoengineering come out on different sides for different countries.

Consider in particular the case with a 2°C change in temperature, and a 1% of GDP

[21] We could formalize this by assuming that there is a distribution of dangers of geoengineering that are multiplicative of GDP, much as the damages functions in the RICE model are, and that α represents the certainty equivalence of these dangers. Such a model would generate region specific α_k's, but given our lack of understanding of these dangers we will restrict our attention to a uniform α for simplicity.

[22] The natural comparison is therefore to Table 2.3. Considering current GDP does not affect which countries are for or against geoengineering in our scenario, because both sets of damages are multiplicative. It does, however, skew our analysis in favor of believing a negotiated outcome is possible, because regions that are against geoengineering in our results (see below) are in general currently much richer than nations that would prefer its use.

cost associated with the risks of geoengineering. Then Africa, India, other non-OECD Asia and the Middle East will seek to geoengineer, while the US, EU, Russia, Japan, other High Income countries, Latin America and China would all prefer restraint. In aggregate, the world loses from geoengineering in this setting (the aggregate net losses are US$148 billion - not enormous, but certainly non-trivial). We then have a world in which a small set of countries will pursue geoengineering at the expense of the global community.

This example also points to the possibility of a negotiated solution to this problem, provided international agreements about geoengineering are enforceable. In a 2°C temperature change, 1% of GDP risk-associated cost of geoengineering world, the potential damages from geoengineering to the US alone are larger than the total gains to all of the countries that would seek to free-drive. It is therefore possible that in such a scenario the US alone could convince those countries not to geoengineer, in return for aid or adaptation assistance.

It is worth comparing these results to the few examples we have of attempts at a cost-benefit analysis at a global level. Goes *et al.* (2011) compare optimal mitigation behavior with geoengineering in the absence of mitigation, assuming that geoengineering has potential damages that are a fixed multiple of global output (similar to our approach here).[23] They suggest that while there is substantial uncertainty, under their assumptions geoengineering is a worthwhile option to pursue only if the costs are approximately 0.5% of Gross World Product or less.[24] Taking a different approach that focuses on sea-level rise, Moore *et al.* (2010) suggests that the cost-benefit threshold for geoengineering is approximately 0.6% of GWP. While they are generated in quite different settings, these global thresholds are remarkably close to the estimates for the risks of geoengineering at a regional level that are

[23]Goes *et al.* (2011) also analyze the possibility of geoengineering being abandoned after 50 years, and the rapid climate change that would result.

[24]Bickel and Agrawal (2011) question these results, arguing that the right comparison is with geoengineering as part of a portfolio of options, including mitigation, and in that case geoengineering is worthwhile under a much wider range of cost assumptions.

likely to lead to free-driving situations in our model. It seems likely then that in cases in which the usefulness of geoengineering is disputed on a global level, that a free-driving dynamic will also be a very real possibility.

Are we likely to see such a dynamic develop in practice? To answer that question, we have to turn to questions of international architecture and governance, and answer tricky questions about enforceability and the free-riding nature of such an agreement. There are a number of global areas in which the US could individually bankroll globally beneficial policy changes, but either chooses not to or isn't given the opportunity for very real political and economic reasons, and geoengineering may be similar. We turn to those question in the final section.

The bottom line is that free-driving is a possible scenario in the context of geoengineering, given the likely degree of heterogeneity in potential damages from climate change. Using the RICE model, a widely accepted integrated assessment model, we have demonstrated the significant diversity of damages from climate change that countries face, and so the heterogeneity in benefits from pursuing technologies that might ameliorate or totally reverse global temperature changes. While the RICE model does not generate disagreement over whether climate change should be avoided, the addition of a cost associated with using geoengineering to reverse its effects does generate such disagreement, and points to the fact that absent an international agreement in this area we could very easily see a free-driver dynamic develop.

2.5 Conclusion and Implications

2.5.1 Theory

Our simple model of public gob provision suggests that, in certain situations, the problem of public good provision may exhibit properties of free-driving as opposed to free-riding. These are situations in which the cost of provision of the public good (gob) is sufficiently low for one actor (or a small coalition of actors) to provide it, and the preferences toward the public good (gob) are sufficiently heterogeneous (such that high levels of the good yield positive utility for some, and negative utility for others). While we have focused on the case of climate geoengineering, other applications of this basic framework abound. Many issues involving new technology are likely to behave in a similar fashion, while the conservation of an endangered species and the introduction of non-native species into a native environment, for example, may also feature a free-driver dynamic. Nuclear proliferation is a possibly more harrowing example of free-driving by "rogue" nations or actors, with adverse welfare consequences for the world as a whole.

One of the key novel insights from our highly simplified model is that inequality and heterogeneity of preferences matters not only for distributional equity but also potentially for economic efficiency. Unlike in the case of pure public goods where heterogeneity in preferences over potential contributors may improve social welfare, in a free-driving world, greater heterogeneity can generate larger discrepancies between the non-cooperative Nash equilibrium and the cooperative social optimum.[25] Reducing this heterogeneity - through in-kind transfers for climate adaptation, for example - may in principle move us away from the free-driving scenario at relatively low cost.

We see the primary value of the model as formalizing the intuition of free-driving, which has been noted in passing by Victor (2008) and others, and suggested as a possibly more

[25]This point is expanded upon in both Appendix B.1 and B.2

general phenomenon by Weitzman (forthcoming). However, we also believe the approach taken here contributes to the theory of public goods provision as built by Samuelson (1954), Bergstrom *et al.* (1986) and others. Appendix B.2 discusses this element of our contribution in more detail.

2.5.2 Policy Implications

This paper has spelled out the implications of world in which countries are capable of pursuing unilateral geoengineering. However, this is only one of several possibilities, and the alternatives are worth considering. The first possibility is that even absent a negotiated agreement, countries will not unilaterally geoengineer for fear of retaliation by other countries. Analogously to the literature on sovereign debt defaults, it is possible that the existence of other dimensions of foreign policy realistically restricts the set of options for countries in the geoengineering sphere. However, retaliation of that nature relies on the ability to detect ongoing geoengineering, and assign blame to the party responsible. The question of feasible monitoring of geoengineering remains an open one in the literature, with David Keith suspecting that it would be very difficult with current technology (see also Robock *et al.* (2010)).[26] If this is the case, then free-driving geoengineering is a very realistic probability.

It is more likely however that eventually an international governance structure for geoengineering will be developed, and countries will acquiesce to the negotiated rules. While there is some suggestion that existing international frameworks, in particular the Convention on Biological Diversity, may be extended to cover geoengineering,[27] it seems

[26] A related possibility is one of "counter-geoengineering", in which countries with preferences for higher temperatures introduce particles of a different nature in a bid to undo some of the cooling effects of SRM. If geoengineering is reversible in this sense, then the interactions are significantly different. David Keith and Andrew Parker are working on a paper of this type.

[27] In 2010, the Convention on Biological Diversity was agreed as banning the use of iron fertilization of the ocean (Secretariat of the Convention on Biological Diversity, 2012).

more likely that a standalone convention or negotiations as part of a new round of climate change talks more generally will be required.

In negotiating such an international agreement there are two prominent issues nations would seek to cover. The first is the circumstances under which full-scale geoengineering could be carried out. Using the model developed in this paper, such a negotiation would seek to restrict action in the free-driving parameter spaces, subject to a participation constraint (it is often said of the Nuclear Non-Proliferation Treaty that it has a very high compliance rate, except for the countries developing nuclear weapons). To ensure that all of the potential free-drivers join such an arrangement, terms would have to be offered to make compliance a better option than remaining outside the treaty framework.

The insight that the framework above generates is that it is possible that transfers and assistance specifically aimed at reducing potential climate damages are more useful than straight monetary transfers. The reason for this is that mitigation and adaption assistance will, in the context of our model, lower the damages from climate change (D_i), and so lower the desirability of geoengineering. Monetary transfers, on the other hand, will likely lower the perceived financial cost of geoengineering (C_i), and so may make free-driving *more* likely unless conditions can be effectively placed on the aid. The difference between directly addressing preference heterogeneity, and addressing income inequality, is stark in the model developed here, and has implications for the treaty design.[28]

The second relevant dimension of treaty design is the governance of research into geoengineering. While the model above is intentionally static, lying behind it is a dynamic game of technology development, and calls have already been made for a moratorium on research and experiments until more is known. Victor (2008) recommends designing

[28]The Nuclear Non-Proliferation Treaty may be a model in this context, offering linked civilian assistance (adaptation technology) in return for forsaking military and internationally dangerous technology (geoengineering).

institutional norms of transparency in geoengineering research, noting that "a taboo would be most dangerous, as it would leave less responsible governments and individuals - those most prone to ignore or avoid inconvenient international norms - to control the technology's fate. A much better approach would be an active geoengineering research program, possible including trial deployments, that is highly transparent and engages a wide range of countries that might have (or seek) geoengineering capabilities." A similar, but more more measured, approach is suggested by Keith and Parsons (2013).

2.5.3 Future Research

A key future research agenda that the topic of this paper prompts is to learn more about heterogeneity in the effects of climate change, and how that affects the thinking of national and regional actors. As Section 2.3 notes, there is a great deal more to learn about differences in how climate change will affect the relevant actors in the modern world.

The model in this paper also needs to be extended to include regional variation in the effectiveness of geoengineering, and its effects on precipitation. Most models suggest that SRM does not work in exactly the opposite direction to the effect of greenhouse gas related climate change, instead having different effects at different levels of latitude and having markedly different effects on precipitation patterns. As a result, the benefits of geoengineering will not simply be equal to the damages from climate change, but the region-specific sum of the effects of global warming related changes and geoengineering related changes. Carrying out such an exercise will give a significantly more accurate characterization of the potential for geoengineering related free-driving.

The second core question is the risks and side effects of different forms of geoengineering, with SRM as a lead example. We currently know very little about the potential effects of planetary scale geoengineering, and without adequate information making decisions in this

area will be fraught with danger (see Keith *et al.* (2010)). We will also need to learn about regional variation in those risks and side effects, as this will affect which nations are, and are not, likely to pursue geoengineering technology.

A related topic, in which economic theory has a role to play as well as empirical work, is to consider the interaction between the geoengineering free-driver problem and the climate change mitigation free-rider problem. One of the main concerns raised about the potential of geoengineering is the concern that technology of that nature will encourage nations to be less vigilant in seeking to prevent climate change in the first place. However, it is also possible that the risk of a single nation taking matters into their own hands and pursuing geoengineering if insufficient effort is made to prevent climate change will act as a spur to ensure that possibility does not arise. A model of this dynamic with two players by Moreno-Cruz (2010) suggests that both effects are possible, and that which effect dominates is dependent on the level of the asymmetry between the two players. More research on this topic is clearly needed.

Some scientists argue that we have entered the era of the anthropocene, where human economic activity has become a dominant geological force affecting the most fundamental of planetary processes. Climate geoengineering represents perhaps an extreme leap in that direction; it has the potential to create an artificial global thermostat, albeit one that features questionable precision and some risk of throwing the boiler violently out of control. A crucial international governance challenge will be to determine: whose hands will be allowed on the thermostat?

Table 2.1: *Summary of Literature*

Study	IPCC (1996)	Mendelsohn *et al.* (2000)	Nordhaus and Boyer (2000)	Tol (2002)	Hope (2006)[a]	Maddison and Rehdanz (2011)[b]	Bosello et al (2012)
Level of Warming	2.5°C	2.5°C	2.5°C	1°C		3.2°C	
North America			3.4 (1.2)				
- United States		-0.5			-4.4 (-0.3,-15.7)	0.3	0.2
OECD Europe				3.7 (2.2)			
- EU			-2.8		-3.9 (-0.3,-13.9)	6.5	0.0
OECD Pacific				1.0 (1.1)			
- Japan						1.0	
Eastern Europe/FSU				2.0 (3.8)	-1.7 (-0.1,-6.1)	-13.4	0.6
- Eastern Europe			0.7				
- Russia		11.1	0.7				
Middle East			-2.0	1.1(2.2)		-12.2	-0.8
Latin America				-0.1(0.6)	-1.8 (-0.1,-6.5)	-40.7	-0.7
- Brazil							
South, Southeast Asia			-1.7(1.1)			-25.4	-3.1
- India		-2.0	-4.9		-9.7 (-1.3,-32)		
China		1.8	-0.2	2.1(5.0)	-3.1 (-0.4,-10.8)	4.5	0.2
Africa			-3.9	-4.1(2.2)	-7.7 (-0.9,-25.3)	-65.1	-4.2
Developed countries	-1.0 to -1.5	0.03					
Developing countries	-2.0 to -9.0	-0.17					

This table is based on a similar table in Smith et al. (2001).

[a] *Calculations done by the authors using the PAGE07 model. Damages are over the full time horizon to 2100, as a discounted fraction of regional GDP. The figures in parentheses represent the 5% and 95% ranges of the simulations. Our thanks to Chris Hope for providing the model and generous assistance.*

[b] *Regional results aggregated by the authors. Maddison and Rehdanz (2011) base their estimates on willingness to accept the impacts of climate change on household welfare. These estimate are, by design, larger than willingness to pay to avoid those impacts.*

Table 2.2: ***RICE 2010 Damage Parameters***

Region	Coefficient on Temperature (ψ_1^k)	Coefficient on Temperature Squared (ψ_2^k)
US	0.0	0.1414
EU	0.0	0.1591
Japan	0.0	0.1617
Russia	0.0	0.1151
Eurasia	0.0	0.1305
China	0.0785	0.1259
India	0.4385	0.1689
Middle East	0.2780	0.1586
Africa	0.3410	0.1983
Latin America	0.0609	0.1345
OHI	0.0	0.1564
Other non-OECD Asia	0.1755	0.1734

Table 2.3: *Damages using 2005 GDP Figures*

Region	Outcome	Temperature Change							
		0	0.5	1	1.5	2	2.5	3	4
US	GDP	12.3979	12.3935	12.3804	12.3585	12.3278	12.2883	12.2401	12.1174
	Utility	0.6905	0.6905	0.6903	0.6900	0.6897	0.6892	0.6885	0.6870
EU	GDP	13.0311	13.0259	13.0103	12.9844	12.9481	12.9015	12.8445	12.6993
	Utility	0.6121	0.6121	0.6118	0.6114	0.6109	0.6102	0.6093	0.6071
hline Japan	GDP	3.8703	3.8687	3.8640	3.8562	3.8452	3.8312	3.8140	3.7701
	Utility	0.6366	0.6365	0.6363	0.6359	0.6354	0.6348	0.6339	0.6318
Russia	GDP	1.6980	1.6975	1.6960	1.6936	1.6901	1.6857	1.6804	1.6667
	Utility	0.4193	0.4192	0.4190	0.4185	0.4179	0.4172	0.4163	0.4139
Eurasia	GDP	0.8073	0.8071	0.8063	0.8050	0.8031	0.8008	0.7979	0.7905
	Utility	0.1210	0.1209	0.1204	0.1197	0.1187	0.1174	0.1158	0.1117
China	GDP	5.3332	5.3295	5.3223	5.3118	5.2980	5.2808	5.2603	5.2091
	Utility	0.0109	0.0105	0.0099	0.0089	0.0076	0.0060	0.0040	-0.0009
India	GDP	2.4408	2.4344	2.4260	2.4155	2.4029	2.3883	2.3716	2.3321
	Utility	-0.3393	-0.3411	-0.3434	-0.3463	-0.3498	-0.3540	-0.3587	-0.3702
Middle East	GDP	3.4801	3.4739	3.4649	3.4532	3.4387	3.4214	3.4014	3.3531
	Utility	0.3112	0.3106	0.3097	0.3085	0.3071	0.3053	0.3033	0.2983
Africa	GDP	1.3005	1.2977	1.2935	1.2881	1.2813	1.2733	1.2640	1.2415
	Utility	-0.5324	-0.5341	-0.5366	-0.5398	-0.5438	-0.5487	-0.5544	-0.5684
Latin America	GDP	4.5585	4.5556	4.5496	4.5405	4.5284	4.5132	4.4950	4.4492
	Utility	0.3019	0.3017	0.3012	0.3005	0.2996	0.2984	0.2970	0.2934
OHI	GDP	3.8420	3.8405	3.8360	3.8285	3.8180	3.8045	3.7880	3.7459
	Utility	0.6333	0.6332	0.6330	0.6326	0.6321	0.6315	0.6307	0.6286
Other non-OECD Asia	GDP	2.6192	2.6158	2.6100	2.6021	2.5918	2.5793	2.5645	2.5281
	Utility	-0.1964	-0.1971	-0.1985	-0.2003	-0.2027	-0.2056	-0.2090	-0.2177

Table 2.4: *The Effect of Side Effects on the Desirability of Geoengineering*

		Geoengineering Risk as a Fraction of GDP				
Region	Δ Temp	0	0.5%	1%	2.5%	5%
US	0	0.0	-62.0	-124.0	-309.9	-619.9
US	2	70.1	8.1	-53.9	-239.8	-549.8
US	4	280.5	218.5	156.5	-29.4	-339.4
EU	0	0.0	-65.2	-130.3	-325.8	-651.6
EU	2	82.9	17.8	-47.4	-242.8	-568.6
EU	4	331.7	266.6	201.4	6.0	-319.8
Japan	0	0.0	-19.4	-38.7	-96.8	-193.5
Japan	2	25.0	5.7	-13.7	-71.7	-168.5
Japan	4	100.1	80.8	61.4	3.4	-93.4
Russia	0	0.0	-8.5	-17.0	-42.4	-84.9
Russia	2	7.8	-0.7	-9.2	-34.6	-77.1
Russia	4	31.3	22.8	14.3	-11.2	-53.6
Eurasia	0	0.0	-4.0	-8.1	-20.2	-40.4
Eurasia	2	4.2	0.2	-3.9	-16.0	-36.2
Eurasia	4	16.9	12.8	8.8	-3.3	-23.5
China	0	0.0	-26.7	-53.3	-133.3	-266.7
China	2	35.2	8.6	-18.1	-98.1	-231.4
China	4	124.2	97.5	70.8	-9.2	-142.5
India	0	0.0	-12.2	-24.4	-61.0	-122.0
India	2	37.9	25.7	13.5	-23.1	-84.1
India	4	108.8	96.6	84.4	47.7	-13.3
Middle East	0	0.0	-17.4	-34.8	-87.0	-174.0
Middle East	2	41.4	24.0	6.6	-45.6	-132.6
Middle East	4	127.0	109.6	92.2	40.0	-47.0
Africa	0	0.0	-6.5	-13.0	-32.5	-65.0
Africa	2	19.2	12.7	6.2	-13.3	-45.8
Africa	4	59.0	52.5	46.0	26.5	-6.0
Latin America	0	0.0	-22.8	-45.6	-114.0	-227.9
Latin America	2	30.1	7.3	-15.5	-83.9	-197.8
Latin America	4	109.2	96.4	63.6	-4.7	-118.7
OHI	0	0.0	-19.2	-38.4	-96.1	-192.1
OHI	2	24.0	4.8	-14.4	-72.0	-168.1
OHI	4	96.1	76.9	57.7	0.1	-96.0
Other non-OECD Asia	0	0.0	-13.1	-26.2	-65.5	-131.0
Other non-OECD Asia	2	27.4	14.3	1.2	-38.1	-103.6
Other non-OECD Asia	4	91.1	88.0	74.9	25.6	-39.9

The values in each cell are the difference between GDP with geoengineering, including some fraction taken out as a certainty equivalent of the risks of side effects, and GDP with climate change damages. Green cells contain positive numbers (the nation prefers geoengineering), red cells contain negative numbers (the nation prefers restraint). All calculations use 2005 GDP Figures in billions of US$.

Chapter 3: A Citizen-Candidate Model of Primary Elections

3.1 Introduction

Two common explanations for policy divergence between political parties are policy motivated candidates and internal party dynamics, in particular primary elections. The first explanation argues that divergence comes from a trade-off between electability and policy outcomes, while the second focuses on a trade-off between winning the primary election and winning the general election.

In this paper we combine these two approaches to explain not only policy policy divergence between parties, but also divergence within primary elections. In doing so, we also investigate the behavior of actors in both types of model, and establish conditions for these models to be well-behaved.

The model we will be working with has two main phases: a general election and a primary election. The general election features two policy-motivated candidates in the style of Wittman (1983) and Calvert (1985), and is the focus of Section 3.2. The primary elections take the form of a citizen-candidate model in which every member of a political party is potentially a candidate for office. This phase is most closely related to Osborne and Slivinski (1996), and is the focus of Section 3.3. Section 3.4 characterizes the equilibria of

the combined model, considering elections in a party facing an incumbent opponent for simplicity, and contains the key results. Section 3.5 concludes discusses extensions, future research and concludes. The remainder of this section discusses the relationship of this paper to the existing literature and the key results.

Relationship to the literature

The idea that policy divergence between political parties is an anomaly that requires explaining dates back to Hotelling (1929) and Downs (1957). These models predict policy convergence in two candidate systems, as the two candidates move their proposed platforms closer to each other in a bid to win over the median voter. While we do observe parties with similar platforms in many elections, we rarely, if ever, see complete policy convergence, and quite often observe parties putting forward radically different platforms.

Through the 1970s and 1980s, a number of authors suggested that this counter-factual result was driven by Downs's assumption that parties and candidates only care about winning elections, and not about the policies that are implemented. If candidates instead care about both winning and policy outcomes, and there is some uncertainty about the position of the median voter, then the two platforms will diverge in equilibrium. This argument was made by Wittman (1977) and Calvert (1985), amongst others, and is a key part of the general election phase of the model in this paper.

While this approach resolves the question of where policy divergence comes from, it relies on exogenously given policy positions for the candidates. This generates a new question: where do candidate preferences come from? The natural answer to this question is that they are generated in equilibrium by either competitive elections or negotiation within the party itself, which leads us to want a model of this process.

Looking at the party primary process as part of the explanation for policy divergence dates to roughly the same period as the papers using policy-motivated candidates: see for example Coleman (1971) and Aranson and Ordeshook (1972).[1] Where candidates must first win over the members of their own party, who are presumably more partisan than voters in the general election, and candidates cannot casually change platforms or preferences between the primary election and the general election, then we will again observe policy divergence. Recent papers that take this approach include Jackson *et al.* (2007), Owen and Grofman (2006) and Chen (2009).

However, this approach, and each of the above papers, predicts policy convergence *within the party*: the winner of the primary is always the preferred candidate of the party median. Since we observe substantial policy divergence in primary elections as well as in general elections, more is needed. The model in this paper generates that diversity within a competitive primary, and nests the possibility that the winning candidate is the preferred candidate as the party median as one version of the one-candidate equilibria.[2] However, for most types of equilibria in our model the winning candidate will *not* be the preferred candidate of the party median.

The citizen-candidate model of elections used in this paper was proposed by Osborne and Slivinski (1996).[3] In the Osborne and Slivinski model, a population of citizens has preferences over a one-dimensional set of policies, and chooses whether to become a candidate in the election by paying some fixed cost. The winner of the election implements their preferred policy and gains a benefit from simply holding office.[4] This model generates dispersion in platforms by the candidates contesting an election. Osborne and Slivinski use

[1] See also Hansson and Stuart (1984) and Aldrich and McGinnis (1989).

[2] To foreshadow the details of the model, if $b = c = 0$, then we can obtain the equilibria in Owen and Grofman (2006) and Chen (2009).

[3] Besley and Coate (1997) independently developed a similar model.

[4] Using the terminology of Rogoff (1990), Osborne and Slivinski (1996) refer to these as "ego rents".

this model to characterize different equilibria (similar to the approach taken in this paper) and to compare the results under plurality and runoff voting rules.

The approach taken in this paper adresses the gaps in each of the approaches described above. It explains where the policy positions of the general election candidates come from (the policy preferences of the equilibrium winners of the primary elections), while also generating a range of equilibria in which there are multiple potential winners of the primary elections, each with different preferred platforms. Finally, it sets the citizen-candidate model in a context in which the assumption that every voter is a potential candidate seems much more plausible - the party primary process.

The idea of modeling primary elections using a citizen-candidate framework was first proposed by Cadigan and Janeba (2002). However, because their model assumes a framework of certainty in both primary and general elections, it can only generate policy divergence within parties by assuming voters naively vote for the candidate closest to their position. When party members instead vote for the candidate that will give them the highest expected utility, there is no differentiation of candidates within parties, with typically only the most moderate member of the party the only viable candidate. By contrast, this paper generates policy divergence both within and between parties by including uncertainty over the outcome of the general election.

A related question, that this paper does not touch on, is why political parties hold primaries. The dominant suggestion offered in the literature so far is that primaries are used to reveal the quality of potential general candidates (see Serra (2011) and Adams and Merrill (2008)). We do not consider candidate valence issues in this paper, focusing instead on policy dimensions. See also Holden and Hummel (2011) on optimal primary processes.

Key Results

The key result of this paper is that the number of candidates in a primary election depends on the costs of entering and the potential benefit of winning the general election, and that as the number of candidates grows the level of policy divergence within the party tends to grow as well. In particular, as the benefits of winning the general election exceed a certain level, the winner of the primary will generically *not* be the ideal candidate of the party median.[5] Perhaps the most striking of these equilibria is the case in which not only do candidates with platforms more extreme that their party median run, but they can also win both the primary election and (given probabilistic voting) the general election.[6] These results are discussed in Section 3.4.

The other results concern the behavior of both the general election and primary election models. In the general election, we are most interested in how changes in the policy preferences of the candidates flow into changes in their announced policies. Surprisingly, this comparative static does not appear to have been addressed in the literature, and is ambiguously signed without further assumptions. We characterize those assumptions in Section 3.2. In the primary election, for us to easily characterize equilibria we require that voter preferences obey the single-crossing property. Remarkably, the assumption that guarantees single-crossing is exactly the same assumption that enables us to give the relationship between general election candidate preferences and platforms its intuitive sign. This is discussed in Section 3.3.

Readers should note that while this paper uses the language of primaries, referring most directly to the American electoral system, the model of this paper applies equally

[5]The notion that the party median often gets "squeezed out" by extremists on either side is discussed in Brams (1978).

[6]Simulation results with exogenously given candidate positions in the primary generates similar results - see Cooper and Munger (2000).

wherever parties elect a leader to contest a general election. The effects analyzed here play themselves out in leadership contests in parliamentary democracies just as much as in presidential democracies if we abstract away from the issue of endogenous timing of leadership challenges. Indeed, because in parliamentary leadership challenges every voter is, in a very real sense, a potential candidate, they form an even more natural setting for the citizen-candidate model.

3.2 General Election

We begin by focusing on the behavior of candidates in the general election. The idea of candidates with policy preferences as a potential resolution of the counter-factual full convergence predictions of the traditional Downs (1957) model has a long heritage, and can be traced back at least as far as Wittman (1977), with much of the argument foreshadowed in Wittman (1973).[7] Further developments were also made by Hansson and Stuart (1984) and Calvert (1985). These models generate a trade-offs for candidates between offering a policy close to their own preferences and electability, which in the presence of uncertainty about voter behavior generates partial, but not complete, convergence.[8]

This model of general elections is a building block of our later full model. We develop it in full here for that reason, and also because there is a previously unexplored comparative static that is related to a result we will need in the full model.

---

[7]Roemer (2006) refers to policy motivated candidates as the "Wittman model".

[8]The necessity of both policy-driven candidates *and* voter uncertainty is recognized in Wittman (1983) and emphasised by Roemer (1997).

3.2.1 Model

Assume two parties, Left (L) and Right (R), each with a candidate. Each candidate i has preferences $u_i(w, a_i)$ over the policy which is implemented, w, given their preferred policy position a_i. We take these policy preferences as given for now. We assume $u_i(w, a_i)$ is a decreasing concave function of the distance between w and a_i. Candidates also receive some benefit from winning the election b.

In the general election between these two candidates, each candidate can credibly commit to a policy platform x_i.[9] Given announced positions x_L and x_R, we assume the existence of a function $P(x_L, x_R)$ that describes the probability of the Left candidate winning the election. We assume $P(x_L, x_R)$ is increasing in x_L for $x_L < x_R$, decreasing for $x_L > x_R$, and continuous in x_L and x_R except possibly at $x_L = x_R$.

This function can be motivated in a number of ways (see Roemer (2006)).[10] One natural way to derive the function is to assume that each voter votes for the candidate that delivers them the highest utility, but there is uncertainty about the distribution of the voters' policy preferences.

Under these assumptions, the Left candidate's problem in choosing a policy platform x_L, given some announced policy by the Right candidate of x_R, is:

$$\max_{x_L} \; P(x_L, x_R)\,[b + u(x_L, a_L)] + (1 - P(x_L, x_R))u(x_R, a_L)$$

[9] We follow the literature in assuming candidates can credibly commit to policies. This assumption is not without consequences or alternatives (see Alesina (1988)). We comment on this assumption further in Section 3.3.

[10] Depending on the choice of microfoundations, this function is not necessarily well-behaved (see Roemer (1997) and Roemer (2006)). We assume that the function is well-behaved, at the potential expense of linking the results to model primitives.

This is the familiar expression of the problem. Wittman (1983) and Calvert (1985) show that in this model, there exists a Nash equilibrium in policy announcements by the two candidates. Moreover, in general $a_L < x_L < x_R < a_R$. That is, each candidate announces a platform more moderate than their preferred policy, but different from their opponent's.

3.2.2 The effect of policy preferences on platforms

Let us now consider the candidates' trade-offs in more detail. The first-order condition of the candidate's problem is:

$$\underbrace{\frac{\partial P(x_L, x_R)}{\partial x_L}}_{+ve} \underbrace{[b + u(x_L, a_L) - u(x_R, a_L)]}_{+ve} + \underbrace{P(x_L, x_R)}_{+ve} \underbrace{\frac{\partial u(x_L, a_L)}{\partial x_L}}_{-ve} = 0$$

The candidate's decision is determined by two forces. The first is the fact that becoming more moderate makes victory more likely, and the candidate prefers victory over defeat, both because they prefer to be in office (b) and because they prefer their policy over their opponent's ($u(x_L, a_L) - u(x_R, a_L)$). This leads candidates to become more moderate.

On the other hand, becoming more moderate leads candidates to announce a policy that they prefer less than positions closer to their bliss point. In equilibrium, candidates never nominate a policy more extreme than their own preferences, and so in any equilibrium $\frac{\partial u(x_L, a_L)}{\partial x_L}$ is negative. This forces leads candidates to become more extreme. Equilibrium is pinned down by the equation of these two forces.

The comparative statics of this equilibrium have been investigated from the start of this literature. Wittman (1983) and Calvert (1985) discuss the effect of changes in the probability function, while Roemer (1997) gives the most thorough treatment and shows that increases in uncertainty push the candidates towards their preferred policies. Alesina (1988) notes that the distance between the two policies is inversely related to the benefits of gaining

office b. More recently, Owen and Grofman (2006) use specific functional forms to consider comparative statics in a series of examples, while Mirhosseini (2007) shows that in symmetric equilibria more extreme incumbents, less favorable electorates and greater certainty about the probability function lead to more moderate platforms, but that in asymmetric equilibria the effects may be non-monotonic.

However, one comparative static appears to have not been addressed in the literature: how does moving the policy preference of the candidate affect their choice of platform? The answer is that it is unclear! While it seems intuitive that more moderate candidates select more moderate platforms, there are in fact again two forces here, and which one dominates determines whether platforms become more moderate or more extreme as candidates' preferences become more moderate .

To see this, we use the implicit function theorem to define x_L^* as a function of the parameters (including, for now, x_R, which we are taking as given), and then differentiate with respect to a_L. Re-arranging, we obtain:

$$\frac{\partial x_L^*(a_L)}{\partial a_L} = \frac{\frac{\partial P(x_L,x_R)}{\partial x_L}\left[\frac{\partial u(x_L,a_L)}{\partial a_L} - \frac{\partial u(x_R,a_L)}{\partial a_L}\right] + P(x_L,x_R)\frac{\partial^2 u(x_L,a_L)}{\partial x_L \partial a_L}}{-\left[\frac{\partial^2 P(x_L,x_R)}{\partial x_L^2}\left[b + u(x_L,a_L) - u(x_R,a_L)\right] + P(x_L,x_R)\frac{\partial^2 u(x_L,a_L)}{\partial x_L^2} + 2\frac{\partial P(x_L,x_R)}{\partial x_L}\frac{\partial u(x_L,a_L)}{\partial x_L}\right]}$$

Concavity of the utility and probability functions guarantee that the the denominator is positive.[11] We can therefore focus on the two terms of the numerator.

[11] This is a part of the proof in Calvert (1985). If the probability function is not concave, then we assume that the concavity of the utility function is sufficient to guarantee that the denominator is positive in the relevant ranges.

$$\underbrace{\frac{\partial P(x_L, x_R)}{\partial x_L}}_{+ve}\underbrace{\left[\frac{\partial u(x_L, a_L)}{\partial a_L} - \frac{\partial u(x_R, a_L)}{\partial a_L}\right]}_{-ve} + \underbrace{P(x_L, x_R)}_{+ve}\underbrace{\frac{\partial^2 u(x_L, a_L)}{\partial x_L \partial a_L}}_{+ve}$$

Begin with the second term. It reflects the fact that as the candidate becomes more moderate in their preferences, a shift in announced policy towards the center is less costly, because the utility function is less curved for a more moderate candidate. This force makes more moderate candidates offer more moderate policies.

Now consider the first term. This term captures the fact that for a candidate more moderate in their preferences, losing the election is less costly, because the opposition's position is less unpalatable. Because losing is less costly, this leads candidates to offer more extreme policies, closer to their own preferences.

Which effect dominates? For the first effect to dominate and $x_L^*(a_L)$ to be increasing in a_L, we require that the numerator be positive. Re-arranging, that requires:

$$\frac{\frac{\partial P(x_L, x_R)}{\partial x_L}}{P(x_L, x_R)} < \frac{\frac{-\partial^2 u(x_L, a_L)}{\partial x_L \partial a_L}}{\frac{\partial u(x_L, a_L)}{\partial a_L} - \frac{\partial u(x_R, a_L)}{\partial a_L}}$$

It is not immediately apparent how and when this condition holds. Essentially, it suggests that for more moderate candidates to offer more moderate policies, we require that the utility function be more curved than the probability function, and so the lower costs of gaining extra probability of winning for moderate candidates outweighs the diminished gap in utilities.

An example of a situation in which this condition holds universally is one in which the $P(x_L, x_R)$ function is derived from a uniform median voter distribution with mean of 0 and bounds $[-\beta.\beta]$, and candidates have quadratic preferences $u(x_L, a_L) = -(x_L - a_L)^2$. In that

case:

$$
\begin{aligned}
P(x_L, x_R) &= \frac{1}{2} + \frac{\frac{x_L + x_R}{2}}{2\beta} \\
\frac{\partial P(x_L, x_R)}{\partial x_L} &= \frac{1}{4\beta} \\
\frac{\partial^2 u(x_L, a_L)}{\partial x_L \partial a_L} &= 2 \\
\frac{\partial u(x_L, a_L)}{\partial a_L} - \frac{\partial u(x_R, a_L)}{\partial a_L} &= 2(x_L - a_L) - 2(x_R - a_L) = 2(x_L - x_R)
\end{aligned}
$$

Substituting in, we require:

$$
\begin{aligned}
\frac{\frac{1}{4\beta}}{\frac{1}{2} + \frac{\frac{x_L + x_R}{2}}{2\beta}} &< \frac{-2}{2(x_L - x_R)} \\
\implies -\beta &< x_R
\end{aligned}
$$

$x_R < -\beta$ is implausible, and the condition for monotonicity of candidates' positions in candidate preferences holds for this example.

More generally, in similar situations it is common to impose that the term on the left, which is the probability distribution's reversed hazard ratio, is decreasing.[12] This is equivalent to stating that the voter preference distribution is log-concave in x_L, a property which holds for a wide range of commonly used distributions, including the uniform and normal. Concave utility, on the other hand, implies that the right hand side is increasing. The combination of these two properties implies that for some finite x_L the right hand side will be greater than the left hand side. Whether that crossing occurs at a plausible value of x_L,

[12]Reversed hazard rates have typically received less attention that hazard rates, but are useful in a range of settings beyond this one. See Veres-Ferrer and Pavía (2012).

however, is an empirical question that depends on the example (although, as we have just seen, it is not difficult to construct an example).

One example that emphatically does *not* generate monotonicity of platforms in preferences is one with that linear utility ($u(x_L, a_L) = -|x_L - a_L|$). With preferences of that sort, all candidates face the same marginal disutility from platform shifts, and given they all face the same probability function will all agree on the same ideal platform to announce. In such a case, the primary process would appear to be completely irrelevant to the general election - a very undesirable result! This is an outcome that does not appear to have been noted in the literature thus far, and motivates our imposition of concave preferences as distinct from the linear preferences used in Osborne and Slivinski (1996), despite the additional complexity this generates.

3.3 Primary Election

We now add in primary elections for each party, in which every member of a party is a potential candidate for election in the primary, and ultimately in the general. It is in this sense that we add a citizen-candidate primary election stage to the familiar Wittman model.

3.3.1 Model

Assume that all citizens are either members of one of the two parties or are independents, and that party affiliation is exogenously determined. The ideal policy positions of the members of the Left party are distributed according to distribution F^L. The median voter of the Left party is denoted by μ_L, while the left most and right most members, if they exist, are L_l and L_r respectively. Similarly, the Right party is distributed according to F^R. It will occasionally be useful to refer to L_l and R_r as the most "extreme" members of their parties and as the "outside" citizens, and L_r and R_l interchangeably as the most moderate, the most centrist and the "inside" citizens.

To determine the candidates in the general election, each party holds a primary election. The electoral process is as follows:

1. **Entry:** Members of the Left and Right parties simultaneously announce whether they intend to enter the primary race (E) or not (N). If no candidate enters, every member of the party receives a payoff of -∞.[13]

2. **Primary Election:** Members of the two parties vote in the primary elections for which candidate will be their nominee for the general election. The winners are determined by plurality rule. If there is a tie the winner is determined at random.

3. **Platform Selection:** The two winning candidates then announce a binding policy platform for the general election, taking into account the electability of their platform and how close it is to their preferred outcome policy outcome.

4. **General Election:** The general election is then determined by all citizens voting. The winner implements their announced policy platform.

The final two stages of this process are the general election process described in Section 3.2. As noted there, these stages have well-defined equilibria, given the outcomes of the earlier stages.

Now consider the primary election. We assume party members vote sincerely[14] for the candidate that maximizes their expected utility, taking into account the uncertainty in the general election, the tendency for their party's candidate to moderate their platform in the general and each of the potential candidates of the other party and their potential chances of winning. For example, the expected utility for a member of the Left party with preferred position a from voting for a candidate with position x_i would be defined as:

[13]This assumption is simply to ensure that at least one candidate enters at the primary stage.

[14]Cf. Besley and Coate (1997).

$$U(x_i, C(R), a) = \sum_{j \in C(R)} \Pi(x_j) \left\{ P(x_L^*(x_i), x_R^*(x_j)) u\left(x_L^*(x_i), a\right) + \left(1 - P(x_L^*(x_i), x_R^*(x_j))\right) u\left(x_R^*(x_j), a\right) \right\}$$

where $C(R)$ represents the set of candidates from R, Π_j is the probability that candidate j with preferred position x_j will win the Right party primary. For simplicity, in Section 3.4 we will assume that the Right party is the incumbent, and x_R is in fact fixed.

Finally, consider the candidate entry stage. There are three incentives for citizens to enter the primary elections. If they win the general election, they receive an ego benefit of b (the same b as in Section 3.2), and they have the right to implement the policy of their choice. They may also seek to influence to identity of the winner of the primary. There is a cost associated with entering the primary campaign of c. Both b and c are non-negative and finite.[15]

If the winner implements policy w, then the payoffs for a citizen with position a are:

$$U(x_i) = \begin{cases} -u\left(x_i, w\right) & \text{if N} \\ -c - u\left(x_i, w\right) & \text{if E, fail to win general election} \\ b - c - u\left(x_i, w\right) & \text{if E, win primary and general} \end{cases}$$

There is no uncertainty about the outcome of the primary elections, except in the case of a tie, and primary candidates cannot pretend that their policy preferences are anything other than what they are.

An equilibrium across all of the stages of the game is a situation in which all citizens are voting for their most preferred candidate in the general and the primary election, the candidates in the general election are playing a Nash equilibrium, all candidates in the primary elections prefer to remain in the election, and all party members who are not

[15]We might also consider the possibility of ego benefits from winning the primary, and of their being some additional cost associated with campaigning for the general after winning the primary, but these do not fundamentally alter the results.

currently candidates would not prefer to stand for office.

3.3.2 Properties of $U(x_L, x_R, a)$

In voting models, it is common to assume one of two properties. The first is single-peaked preferences, which enables the application of the median voter theorem. The second is the single-crossing property, such that if a voter is indifferent between two candidates, all voters to their right prefer the right-most candidate and all voters to the left to their left prefer the left-most candidate. Gans and Smart (1996) and Saporiti and Tohmé (2006) show single-crossing is enough to establish an alternative version of the median voter theorem, provided voters are voting sincerely for the candidate that gives them the highest expected utility (which we assume).[16]

For our results, we will impose both assumptions. We rely more heavily on single-crossing than on single-peakedness, because single-peakedness and the ability to apply the median voter theorem is sufficient for one-candidate equilibria (Propositions 1) but neither neccesary nor sufficient in situations in which the median voter is not necessarily the key actor (the remainder of the results). However, single-peakedness is a useful property, and as we shall see requires relatively little to guarantee.

Our main difficulty in imposing these assumptions is that even if we assume that the utility function over policy outcomes $u(x, a)$ has both of these properties, the induced utility function over primary election candidates does not inherit these properties. This is most easily understood in the context of single-peakedness, and we then turn to single-crossing.

The fact that the induced utility function is not necessarily single-peaked, even if un-

[16]Saporiti and Tohmé (2006) also gives a strategic foundation for a median voter theorem type result under single-crossing, suggesting that we may not need to assume sincere voting for our results. For technical reasons related to the use of a continuum of voters, we continue to make that assumption.

derlying preferences are, comes from a particular form of strategic voting. It is possible that moderate party members can prefer the incumbent over extreme candidates of their own party, and therefore potentially prefer to elect even more extreme candidates of their own party with exceedingly low chances of victory in order to guarantee victory for the incumbent. This problem is noted in Owen and Grofman (2006), Mirhosseini (2007) and Chen (2009).

One approach to avoiding this problem is taken by Owen and Grofman (2006), who assume that it is safe to simply ignore strategic voting of this kind. An alternative approach, used in Chen (2009), makes sufficient assumptions to ensure that this form of strategic voting by moderate party members is not an equilibrium.[17] In particular, Chen (2009) assumes that the composition of each party is such that every member prefers all policies in their own party's platform to any policy from the other party:

Assumption 1 *The set of citizens in each party is such that for any voter a in L, $u(L_l, a) > u(R_l, a)$; and for any voter a in R, $u(R_r, a) > u(L_l, a)$.*

Assumption 1 is not, however, sufficient to guarantee single-peakedness. We further require a decreasing reversed hazard rate in the probability function $P(x_L, x_R)$:

Assumption 2 *The reversed hazard rates of the probability distribution $P(x_L, x_R)$, $\frac{\frac{\partial P(x_L, x_R)}{\partial x_L}}{P(x_L, x_R)}$ and $\frac{\frac{\partial P(x_L, x_R)}{\partial x_R}}{P(x_L, x_R)}$, are decreasing in x_L and x_R respectively.*

As noted in Section 3.2, decreasing reversed hazard rates are equivalent to log-concavity of $P(x_L, x_R)$ and are a natural assumption in this context. Using this assumption, we can demonstrate that the induced utility function is single-peaked:

[17] A further alternative, due to Mirhosseini (2007), uses the fact that under the assumption of quadratic utility, the median voter is still determinative, even if the coalition supporting the median voter's preferred candidate cannot be identified. This approach could be used for Propositions 1 and 2, but is not sufficient for the remainder of the paper. This approach would also restrict us to quadratic utility.

Lemma 1 *If Assumptions 1 and 2 hold, then $U(x_i, x_R, a)$ is single-peaked for $x_i \in [\underline{x}, L_r]$, where $\underline{x}$ is defined by $P(\underline{x}, y) = 0$.*

Obtaining the single-crossing property is more complicated. However, it turns out that a sufficient condition for single-crossing in the induced preferences over primary candidates is precisely the condition in Section 3.2 that guaranteed monotonicity of platforms in candidate preferences, as Lemma 2 states.

Lemma 2 *For $U(k, x_R, a) - U(j, x_R, a)$ to be strictly increasing in $a \in L$ for $k > j$, it is sufficient that $\dfrac{\partial P(j,r)}{\partial j} < \dfrac{-\dfrac{\partial^2 u(j,a)}{\partial a \partial j}}{\dfrac{\partial u(j,a)}{\partial a} - \dfrac{\partial u(r,a)}{\partial a}}$ for all j and a.*

It is an interesting symmetry of the policy-driven candidate and the primary-voting citizen that we require the same relationship between the curvature of the probability function and the curvature of the utility function for their problems to be well-behaved. Intuitively, however, it is perhaps unsurprising: both conditions require that moderate actors are more willing to announce or support more moderate positions than extreme actors. This symmetry also resolves another potential problem: it would be possible for voters to have single-crossing preferences over candidate positions, but for those candidates' announced platforms to fail to obey the same ordering. By guaranteeing that candidates' platforms are monotonically increasing in their positions, the single-crossing property is preserved.

3.3.3 Commentary on assumptions

It is worth commenting on a number of the other assumptions made in this model. First, consider the assumption that candidates can offer binding policy platforms in the general election. As Alesina (1988) points out, there is no *a priori* reason to believe that candidates can be bound to the promises they make during election campaigns in the absence of repeated interaction over time. Our model, on the other hand, relies on the existence of such an ability to commit. A useful metaphor might be to consider the commitment device to be the nomination of a running-mate or potential cabinet. Alternatively, we might imagine

constraints being placed on the candidate by the framework of their own party and the requirement that they work with a Congressional (or similar) mandate.[18]

An alternative approach would be to assume that candidates cannot commit to campaign promises. In this case, candidates are bound to their preferences, and can not offer different policies in the general election. While this is perhaps initially appealing, it leads to the somewhat counter-intuitive situation in the general election that no party-member is their own preferred candidate - every citizen would prefer to nominate a party-member more moderate than themselves, in the process of trading off electability and policy preferences. To avoid this situation, and to (hopefully) keep the intuition clearer, I have pursued the alternative approach.[19]

The ability to commit to a platform in the general election is also in tension with not allowing candidates to commit to any set of policy preferences other than their true preferences in the primary election. This approach is taken in order to keep this paper in line with the literature on citizen-candidate models, and while it is not necessary, removing this assumption would complicate the model dramatically. One way of thinking about this assumption is simply that there is no party or democratic infrastructure to hold individual candidates to their primary promises the way there is in general elections, and that party-members are capable of seeing through promises otherwise.

An alternative approach to studying the impact of primary elections would be to model some cost of flip-flopping between the primary and general elections for candidates who have no policy preferences of their own (see, for example, Hummel (2010)) . However, we

[18]It has been argued that the constraining of candidates for office is one of the roles that parties play in the modern political system. For more, see Aldrich (2011).

[19]There is an interesting analogy here to the literature on delegation of monetary policy to central bankers with more conservative preferences than the average citizen (see, for example, Rogoff (1985)). In a model without commitment, party members would delegate political action to those who can credibly commit to centrism by actually being centrists.

are then struck by the question of how we determine the equilibrium number of candidates in a primary election, and where their preferences come from. Similarly, in discussing the early models in which candidates have preferences over policies as well as winning office, Coleman (1971) mentions a "factional" stage lying behind his model. However, if the factional stage takes the same form as the general election, we have only pushed the question one step further back - where do policy preferences for candidates within factions come from? The approach taken in this paper gives us some insight into the range of candidate equilibria we may have within party primaries, and does so with reference to the characteristics of the general electorate and the party itself.

Finally, we might note that a similar tension exists between the assumption of uncertainty in the general election but full certainty regarding the primary election. Part of the motivation for this assumption is simply that party-members are better informed about the preferences of other party members than they are about the general election, although obviously this is a dramatic simplification.[20] In parliamentary systems, where the relevant "party" is the party caucus, with a small number of very well known primary voters, this assumption is less uncomfortable. In a separate but related paper, Gole (2011) considers the situation where there are two classes of voter - one which knows the results of any potential general election with certainty, and one which is completely uninformed about the general election - and derives results similar to, but more narrow than, the results in this paper.

3.4 Equilibria

I assume that one party is the incumbent, and so their candidate's position is fixed and known in advance as r (to be concrete, I assume the Right party is the incumbent and that it is the Left party that is holding primary elections).

[20] Another part of the motivation is the difficulty of modeling citizen-candidate elections under uncertainty: see Roemer (2004).

We begin by noting the equivalents of Lemmas 1 and 2 from Osborne and Slivinski (1996), which hold analogously for the present model.

Lemma 3 *In equilibrium a candidate does not lose with certainty if either:*

i there are other candidates with the same ideal position as hers; or

ii the ideal positions of all other candidates are on the same side of her ideal position.

Lemma 4 *In any equilibrium at most two candidates share any given position.*

We are now ready to describe the equilibria of our model. We begin with equilibria in which there is only one candidate in the primary election.

3.4.1 One-candidate equilibria

Proposition 1 *One-candidate equilibria facing an incumbent:*

(a) There are one-candidate per party equilibria if and only if $P(x_L^*(\mu_L), r)b \leq 2c$.

(b) If $2c > P(x_L^*(\mu_L), r)b > c$, *then the only one-candidate equilibrium is where the candidate is located at* μ_L.

(c) If $P(x_L^*(\mu_L), r)b < c$, *then a one-candidate equilibrium exists where the candidate* x_c *has a bliss point located in a range defined by the conditions:*

(i) Define $\tilde{x} = x$ *st* $U(x^*(x), r, \mu_L) = U(x^*(x_c), r, \mu_L)$

(ii) $P(x^*(\tilde{x}), y)[u(x^*(\tilde{x}), \tilde{x}) + b - u(y, \tilde{x})] - c - P(x^*(x_c), y)[u(x^*(x_c), \tilde{x}) - u(y, \tilde{x})] < 0$

The results of Proposition 1 generalize Osborne and Slivinski to a party primary setting, and nest the results of Owen and Grofman (2006), Chen (2009) and Mirhosseini (2007). Provided b is not too high, there exists an equilibrium where the median voter inevitably becomes the party's candidate. As b becomes sufficiently high, it becomes worthwhile for another citizen with preferences located at the median to challenge the existing candidate,

with the effect that the pair of them become vulnerable to another candidate entering on either side and winning victory outright. This effect may go some way to explaining why in the absence of incumbents we rarely see single-candidate equilibria in US presidential primary elections, where we would think of the benefits of winning office as very high.

The other type of one-candidate equilibrium is where b becomes sufficiently low that candidates away from the party median can win. This is only an equilibrium if the expected benefit of being the party's nominee is sufficiently low (either because the rewards of office are low or there is little likelihood of victory) that candidates that could successfully challenge the existing candidate would prefer not to. The constraint on how far the candidate can move from the party median comes from the fact that candidates on the other side of the party could enter and, if successful in the general, offer a policy platform they prefer over the existing candidate's.

Looking at the two conditions in (c) more closely, the first defines $\tilde{x}$ as the alternative candidate that the party's median voter is indifferent to, given the existing candidate x_c. The second condition then states that any such candidate must prefer not to enter, as the difference between the expected policy outcome from their entry and the outcome given the current candidate is less than the cost of entry c.

Two further comments are worth making about the equilibrium in (c). The first is that the range of potential candidates will ordinarily not be symmetric around the party median - instead, the constraint will be more binding on the outside of the party than on the inside. This is a result of the fact that when a moderate in the party challenges a slightly-more-extreme-than-the-median candidate, they simultaneously generate the possibility of ego benefits to themselves and increase the party's chance of victory in the general election. A more extreme challenger must trade off the possibility of ego-rents with the decrease in the likelihood of victory in the general, and so is less likely to challenge. This effect speaks

to the fact that parties occasionally nominate very moderate candidates, but rarely very extreme candidates.

The second comment to make is that the bounds on (c) will be wider than the analogous conditions would be in an Osborne and Slivinski setting, because they are factored down by the the likelihood of victory in the general. Indeed, with a sufficiently low probability of victory in the general, the constraints would effectively become non-binding and any candidate in the party would become a potential nominee. All else equal, as the probability of winning increases candidates closer to the party median would find it increasingly desirable to challenge and take the nomination for themselves, restricting the range of equilibria until it converges on the party median itself (the equilibrium in (b)).

3.4.2 Two-candidate equilibria

Define the two candidates in an equilibrium as x_m, the more moderate candidate, and x_e, the more extreme candidate. In stating our result, we will require the following definition: for a set of two candidates x_m and x_e where each of the two is receiving exactly half of the votes, let $s(x_m, x_e, F^L)$ be the location of a challenger who can enter and the two earlier candidates retain equal shares of the vote. Formally, this requires:

$$F^L\{x \text{ st } U(x^*(s), r, x) = U(x^*(x_e), r, x)\} = 1 - F^L\{x \text{ st } U(x^*(s), r, x) = U(x^*(x_m), r, x)\}$$

It is worth noting that even if F_L is single-peaked and symmetric about its median, s does not necessarily equal μ_L.[21] This is because the two newly indifferent people (one between x_e and μ_L, and one between μ and x_m) will not necessarily be symmetric around s.

[21] Cf Osborne and Slivinski (1996).

Our second result is then:

Proposition 2 *Two-candidate equilibria facing an incumbent:*

(a) There are two-candidate per party equilibria if $P(x_L^*(\mu_L), r)b > 2c$.

(b) In any two-candidate equilibrium the candidates' ideal positions are such that the median voter is indifferent: $U(x^*(x_m), r, \mu_L) = U(x^*(x_e), r, \mu_L)$.

(c) A two-candidate equilibrium requires:

(i) The two candidates' positions are sufficiently close together that no there can be no successful challenger between the two candidates: either

$$F^L[x \text{ st } U(x^*(s), r, x) = U(x^*(x_m), r, x)] < 2F^L[x \text{ st } U(x^*(s), r, x) = U(x^*(x_e), r, x)]$$

or

$$F^L[x \text{ st } U(x^*(s), r, x) = U(x^*(x_m), r, x)] = 2F^L[x \text{ st } U(x^*(s), r, x) = U(x^*(x_e), r, x)]$$

and

$$P(x^*(s), y)[u(x^*(s), s) - u(y, s) + b] - \frac{1}{2}P(x^*(x_m), y)[(u(x^*(x_m), s) - u(y, s)] - \frac{1}{2}P(x^*(x_e), y)[(u(x^*(x_e), s) - u(y, s)] \leq 3c$$

(ii) The two candidates' positions are sufficiently far apart that neither finds it optimal to withdraw:

$$P(x^*(x_e), y)b + P(x^*(x_e), y)[u(x^*(x_e), x_e) - u(y, x_e)] - P(x^*(x_m), y)[(u(x^*(x_m), x_e) - u(y, x_e)] \geq 2c$$

(iii) $x_m \neq x_e$.

The results in Proposition 2 again reflect, but are slightly different from, the analogous conditions in Osborne and Slivinski. In any two-candidate equilibrium, the candidates'

positions are neither too identical (lest one withdraw) nor too dispersed (lest there be a successful challenger between them). Note, however, that Part (a) is stated here as an “if" condition, rather than the “if and only if" condition in Osborne and Slivinski. The reason for this is that in that paper they can cleanly identify candidates that keep the median indifferent as being symmetrical around the median, whereas the non-symmetric nature of the probability function here means that we categorically cannot. Even with that constraint, however, Proposition 1’s (a) and Proposition 2’s (a) are collectively exhaustive of the values of b, and as such for any F, c and b there will always be an equilibrium with either one or two candidates.

Part (b) of Proposition 2 notes that in a two-candidate equilibrium, the median voter must always be indifferent between the two candidates. This is because with two candidates there is no reason for a candidate that is surely losing to remain in the race, and so we require both candidates the receive exactly half the votes. This also requires, in contrast to Osborne and Slivinski and along similar lines to Proposition 1, that the more extreme candidate’s position must be more appealing to the median voter than the more moderate candidate’s position, since the extreme candidate is handicapped by their lower chance of winning the general election.

Note that part (c)(ii) of Proposition 2 only refers to the extreme candidate. Because the candidate at x_e selects their optimal policy in the general election, trading off electability and policy, the extreme candidate will always seek to withdraw from the primary election and concede to the other candidate first.

These factors suggest an explanation behind primary races where we observe a centrist candidate with greater general electability facing off against a candidate with greater appeal to the party base, but without losing overall general electability. A recent example of such a race is the 2008 Democratic presidential primary election between Barack Obama and Hillary Clinton, where Clinton was originally cast as more electable, but Obama was more

popular with the base of the party.

3.4.3 Three-candidate equilibria

For the three-candidate equilibrium, define $t_1 = F^{-1}(\frac{1}{3})$ and $t_2 = F^{-1}(\frac{2}{3})$. In a three-candidate equilibrium where all candidates have a chance of winning, we will require that t_1 be indifferent between the first and second candidates from the left, denoted x_1 and x_2, and that t_2 be indifferent between the second and third candidates, x_2 and x_3. This condition, and the formal conditions such that no additional candidate seeks to enter and no existing candidate drops out, are listed in Proposition 3.

Proposition 3 *Every three-candidate equilibrium takes one of the following forms, where the candidates' positions are $x_1 < x_2 < x_3$:*

(a) The positions of the candidates are not all the same, and each candidate receives one-third of the votes. This requires:

(i) the candidates' ideal positions are such that the t_1 and t_2 are indifferent between the two candidates closest to them: $U(x^*(x_1), r, t_1) = U(x^*(x_2), r, t_1)$ and $U(x^*(x_2), r, t_2) = U(x^*(x_3), r, t_2)$.

(ii) The candidates' positions are sufficiently far apart that none find it optimal to withdraw: $P(x^*(x_1), y)[b + u(x^*(x_1), x_1) - u(y, x_1)] - \frac{1}{2}P(x^*(x_2), y)[(u(x^*(x_2), x_1) - u(y, x_1)] - u(y, x_1)] - \frac{1}{2}P(x^*(x_3), y)[(u(x^*(x_3), x_1) - u(y, x_1)] \geq 3c$

(b) The positions of the candidates are all different. Candidates 1 and 3 receive the same fraction of votes, while candidate 2 receives a smaller fraction (and surely loses). This requires:

(i) $P(x^*(x_1), r)b \geq 4c$

(ii) $c < |U(x^*(x_1), r, x_2) - U(x^*(x_3), r, x_2)|$

In the first case of Proposition 3, all three candidates have an equal chance of winning the primary, even though they will naturally have different probabilities of winning the general. As a result, the more extreme candidates must have proposed positions that appeal more to primary voters, in order to make up for a lower chance of victory. The first condition of this case is similar to Proposition 2, in the sense that the candidates must make t_1 and t_2 indifferent. However, there are no restrictions on the distance between the candidates based on the potential for challengers: as long as each candidate receives $\frac{1}{3}$ of the vote, then a challenger would guarantee that the candidate furthest from their own position is the winner. Similarly, there is not a restriction on the minimum distance between candidates. It is entirely possible for two candidates to be located at the position of t_1 or t_2, and to then face off out against a single candidate who will therefore either be a very moderate or very extreme candidate.

This dynamic is evocative of the 2009 leadership contest within the conservative Liberal Party of Australia. Malcolm Turnbull, the leader of the opposition and prominent moderate within the Liberal Party, reached a compromise on a carbon emissions trading scheme with the government, over calls from more conservative elements of his party to reject the scheme. As a result, Mr Turnbull's leadership was challenged, leading to a leadership ballot between Mr Turnbull, Tony Abbot, a former minister associated with the conservative wing of the party, and Joe Hockey, the shadow Treasurer and a moderate, like Mr Turnbull. The party elected Mr Abbot as leader of the opposition, splitting down the middle in electing a more extreme candidate. While the specifics of the story do not quite match the equilibria of Proposition 3 (for example, the leadership ballot was contested by a runoff election, rather than a plurality), they do speak to the intuition behind this result.

The second case of Proposition 3 involves a candidate standing for election with no chance of victory. Instead, they enter in order to draw voters towards themselves, and away from an already standing candidate they do not prefer. In this way, they give a third

candidate, who they do prefer, an equal chance of winning the primary. Osborne and Slivinski (1996) highlights the analogous equilibrium in their model as an interesting, and not necessarily intuitive, outcome of a citizen-candidate model. In the context of early primary elections and caucuses in the US, this dynamic may partially explain the existence of candidates who do not appear to be viable candidates in the longer-run, but who seek to alter the outcome of early race dynamics.

3.4.4 Four or more candidate equilibria

It is very hard to give general conditions for four-candidate equilibria and above. In the basic Osborne and Slivinski model, they can only state some very general restrictions on the types of results that can hold. In our settings, not even an analogous condition holds - we cannot give a necessary condition without imposing concavity on the induced utility function in the primary (as opposed to log-concavity, which is implied by single-peakedness).

Proposition 4 *Four or more candidate per party equilibria: it is possible for an equilibrium in which* $k \geq 3$ *candidates tie for first place to exist even if* $P(x_1, y)b < kc$.

This suggests that the range of four or more candidate equilibria in our model cannot directly be compared to those in Osborne and Slivinski. While Proposition 4 has the ego benefits of victory multiplied by the probability of winning for the most extreme candidate, thus suggesting tighter bounds than Osborne and Slivinski, the absence of concavity of the induced utility function means that our proof cannot proceed even analogously to theirs, and it is possible that the dual trade-offs of policy and electability open up new equilibria that could not exist Osborne and Slivinski. That said, the general intuition from Osborne and Slivinski still holds: as parties become more likely to win the general election, we can expect that their primary elections will feature larger numbers of candidates.

3.5 Discussion

Extensions

There are two natural extensions to the analysis in this paper. The first is to follow Osborne and Slivinski (1996) in comparing the results under plurality rule to results under a runoff system. The second is to consider cases in which both parties are holding primaries. Full exposition of these extensions does not yield great additional insight and so they has been omitted. However, there are some interesting points that can be made, and so we discuss them here.

Comparision with results under a runoff system: Osborne and Slivinski (1996) show that multicandidate elections are less likely under plurality than under a runoff rule and that the maximal dispersion between candidates is smaller under a runoff rule than plurality. These results flow through to our model, with the same forces at work. A runoff system leads to more candidates as challengers can arise at the same position as existing candidates, hoping to win sufficient vote share to enter a runoff they have a chance to win. Similarly, with very dispersed candidates, under a runoff system an entrant need not win outright, but instead must simply acquire sufficient votes to make the runoff, which they can then win.

There is one novel twist: the case in which there are many candidates holding one of two positions is asymmetric not just in positions (see the discussion of Proposition 2) but also in how many candidates can be at each spot. The candidates at the more extreme spot have a lower chance of winning the election and acquiring the ego-rents b, and so fewer candidates can stand there in equilibrium. Since, as in Proposition 9 in Osborne and Slivinski (1996), we require there to be equal numbers of candidates at each of the two clusters (otherwise the top two voter winners are automatically located in the same cluster), this implies that the conditions for such an equilibrium to exist will bind on the extreme cluster before they bind on the more moderate cluster.

Two Parties: Cadigan and Janeba (2002) and Gole (2011) both consider situations in which both parties hold primaries simultaneously. However, the knife edge nature of the results in those papers means that the results are relatively trivial. Under the current model, the results from endogenizing both parties would be non-trivial, but also quite difficult to characterize and an existence proof is not readily available. Assumptions of symmetry between the parties simplify things, but not dramatically. In general, however, it seems likely that the conditions for one party would continue to hold, subject to new statements about how far both parties as a pair could wander from the preferences of their median voters.

With primaries in two parties, equilibrium would require Nash equilibria in both primaries and the general election. The opposing party's selection of candidate changes equilibrium behavior in the general election stage, and so complicates voting in the primary election stage.[22] It would also be possible to compare cases with simultaneous primaries to cases with sequential primaries.[23]

It is, however, clear that the two parties, even if symmetric, can have asymmetric pairs of candidates. This is in contrast to Gole (2011), in which the two parties are forced by the knife-edge nature of equilibria to be symmetric in their primary candidates. It is also worth noting that equilibria in which the parties have differing numbers of candidates are very easy here. Even if both parties are symmetric in all senses, one can easily generate asymmetric equilibria. For example, two candidates from one party (where the party median is indifferent between them) can face a single candidate from the other party. Unfortunately,

[22]Note that this points implies that we do not even have to make the other party's candidate endogenous: we could keep their identity fixed, but just enable them to announce a new policy in the general in response to the winner of the other party's primary. This endogenous policy shift from the opposition would alter equilibrium behavior in the primary. For a foreshadowing of this possibility, see Coleman (1971).

[23]See, by way of comparison, Adams and Merrill (2008).

it seems difficult to give further illumination than the general restrictions above and a sense that "anything goes".

Further Research

There are further political phenomena closely related to the analysis in this paper that are worthy of investigation. The first is open primaries: we have assumed only party members are able to vote in the primary (a "closed" primary), but this is not the only form used by political parties.[24] Open primaries pose a difficulty for modeling, because strategic voting is a first-order consideration,[25] but given their real world prevalence they are a natural topic of future research.

Similarly, primaries have other factors that the simple voting model here does not capture. Two obvious examples are abstention, which has long been proposed as a motivation for policy divergence (see Hinich and Ordeshook (1969)), and contributions. On the topic of contributions, see Alesina and Holden (2008).

The next task would be to locate this model of primary elections within a broader party formation framework. If parties can recruit more moderate or more extreme citizens, then this would potentially dramatically alter the results. This would, of course, require a model of intraparty recruitment decisions, which is a serious issue for research in itself. Recent work along these lines includes Haan (2000), Poutvaara (2003) and Brusco and Roy (2008). Endogenizing the decisions of party members to become active in the party is another interesting related question (see Aldrich (1983)).

Finally, it is worth noting that the approach taken in this paper takes a very narrow

[24] A good summary of the differences between open and closed primaries and their effects is McGhee *et al.* (2013).

[25] See, for example, Chen and Yang (2002) and Oak (2006).

view of what parties are and do, and that there is a wealth of political science literature on the topic of what parties are for and where they come from. This paper is primarily about taking the Downsian/Wittman approach to political outcomes one step further. In reality, this aspect of political parties interacts with their other roles. A good survey of this issue is Aldrich (2011).

Conclusion

This paper develops a model that combined policy-motivated candidates in a general election and citizen-candidates in a primary election. The condition for the two stages of this model to be well behaved is in fact the same condition, and relies on the relative curvature of the probability of victory function and the utility function. Using this model, we developed a set of insights into potential equilibria in primary elections, in which the number of candidates depends on the cost of entry, the benefits of victory and the electability of the party. In general, we observe competitive primaries between candidates with different policy preferences, which in turn explains divergence in the general election. The fact that political parties consistently nominate relatively extreme candidates has long been an interesting question for political economists; in extending the citizen-candidate model to examine the nomination process, I hope that the present work offers some insight into this area.

Bibliography

Adams, J. and Merrill, S. (2008). Candidate and Party Strategies in Two-Stage Elections Beginning with a Primary. *American Journal of Political Science*, **52** (2), 344–359.

Adams, J. B., Mann, M. E. and Ammann, C. M. (2003). Proxy Evidence for an el Nino-like response to volcanic forcing. *Nature*, **426** (6964), 274–278.

Adger, W. N. (2006). Vulnerability. *Global Environmental Change*, **16** (3), 268–281.

Alberini, A., Chiabai, A. and Muehlenbachs, L. (2006). Using Expert Judgment to Assess Adaptive Capacity to Climate Change: Evidence from a Conjoint Choice Survey. *Global Environmental Change*, **16** (2), 123–144.

Aldrich, J. H. (1983). A Downsian Spatial Model with Party Activism. *The American Political Science Review*, **77** (4), 974–990.

— (2011). *Why Parties?: A Second Look*. University of Chicago Press.

— and McGinnis, M. D. (1989). A Model of Party Constraints on Optimal Candidate Positions. *Mathematical and Computer Modelling*, **12** (4), 437–450.

Alesina, A. (1988). Credibility and Policy Convergence in a Two-Party System with Rational Voters. *The American Economic Review*, **78** (4), 796–805.

— and Holden, R. (2008). Ambiguity and Extremism in Elections. NBER Working Paper No. 14143.

Allen, F. and Morris, S. (1998). Game Theory and Finance Applications. In K. Chatterjee and W. Samuelson (eds.), *Game Theory and Business Applications*, Kluwer Academic Publishers, pp. 17–48.

Anwar, S., Bayer, P. and Hjalmarsson, R. (2012). The Impact of Jury Race in Criminal Trials. *The Quarterly Journal of Economics*, **127** (2), 1017–1055.

Aranson, P. H. and Ordeshook, P. C. (1972). Spatial Strategies for Sequential Elections. *Probability Models of Collective Decision Making*.

Arrow, K. (1973). The Theory of Discrimination. In O. Ashenfelter and A. Rees (eds.), *Discrimination in Labor Markets*, Princeton University Press, pp. 3–33.

Ashford, J. and Sowden, R. (1970). Multi-variate Probit Analysis. *Biometrics*, pp. 535–546.

Attanasio, O., Meghir, C. and Santiago, A. (2012). Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA. *The Review of Economic Studies*, **79** (1), 37–66.

Austen-Smith, D. and Banks, J. S. (1996). Information Aggregation, Rationality, and the Condorcet Jury Theorem. *American Political Science Review*, pp. 34–45.

— and Feddersen, T. (2006). Deliberation, Preference Uncertainty, and Voting Rules. *American Political Science Review*, **100** (02), 209–217.

— and — (2009). Information Aggregation and Communication in Committees. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **364** (1518), 763–769.

Bagues, M. and Esteve-Volart, B. (2010). Can Gender Parity Break the Glass Ceiling? Evidence from a Repeated Randomized Experiment. *Review of Economic Studies*, **77** (4), 1301–1328.

Bajari, P., Hong, H., Krainer, J. and Nekipelov, D. (2010). Estimating Static Models of Strategic Interactions. *Journal of Business & Economic Statistics*, **28** (4), 469–482.

Barrett, S. (2007). *Why Cooperate?: The Incentive to Supply Global Public Goods: The Incentive to Supply Global Public Goods*. OUP Oxford.

Baum, S. D., Maher Jr, T. M. and Haqq-Misra, J. (2013). Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse. *Environment Systems & Decisions*, **33** (1), 168–180.

Beck, T., Behr, P. and Madestam, A. (2012). Sex and Credit: Is There a Gender Bias in Lending? *European Banking Center Discussion Paper*, (2012-017).

Bergstrom, T., Blume, L. and Varian, H. (1986). On the Private Provision of Public Goods. *Journal of Public Economics*, **29** (1), 25–49.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review*, **94** (4), 991–1013.

Besley, T. and Coate, S. (1997). An Economic Model of Representative Democracy. *The Quarterly Journal of Economics*, **112** (1).

Bickel, J. E. and Agrawal, S. (2011). Reexamining the Economics of Aerosol Geoengineering. *Climatic Change*, **119** (3-4), 993–1006.

Blanes i Vidal, J. and Leaver, C. (2013). Social Interactions and the Content of Legal Opinions. *Journal of Law, Economics, and Organization*, **29** (1), 78–114.

Boosey, L. (2011). Information, Social Preferences, and Learning in Network Public Goods Experiments. *Working Paper*.

Bosello, F., Eboli, F. and Pierfederici, R. (2012). Assessing the Economic Impacts of Climate Change, fEEM (Fondazione Eni Enrico Mattei), Review of Environment, Energy and Economics (Re3), February 2012. Available at SSRN: http://ssrn.com/abstract=2030223.

Boudreau, K., Brady, T., Ganguli, I., Gaule, P., Guinan, E., Hollenberg, A. and Lakhani, K. (2013). The Formation of Scientific Collaborations: Field Experimental Evidence on Search Frictions in Collaborator Matching. *Working Paper*.

Brams, S. J. (1978). *The Presidential Election Game*. Yale University Press.

Brouwer, R. and Spaninks, F. A. (1999). The Validity of Environmental Benefits Transfer: Further Empirical Testing. *Environmental and Resource Economics*, **14** (1), 95–117.

Brovkin, V., Petoukhov, V., Claussen, M., Bauer, E., Archer, D. and Jaeger, C. (2009). Geoengineering Climate by Stratospheric Sulfur Injections: Earth System Vulnerability to Technological Failure. *Climatic Change*, **92** (3-4), 243–259.

Bruhn, M. and McKenzie, D. (2009). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics*, pp. 200–232.

Brusco, S. and Roy, J. (2008). Aggregate Uncertainty in the Citizen Candidate Model Yields Extremist Parties. *CEDI Discussion Paper Series*.

Cadigan, J. and Janeba, E. (2002). A Citizen-Candidate Model with Sequential Elections. *Journal of Theoretical Politics*, **14** (4).

Calvert, R. L. (1985). Robustness of the Multidimensional Voting Model: Candidate Motivations, Uncertainty, and Convergence. *American Journal of Political Science*, **29** (1), 69–95.

Cameron, A., Gelbach, J. and Miller, D. (2011). Robust Inference with Multiway Clustering. *Journal of Business and Economic Statistics*, **29** (2), 238–249.

Casey, K., Glennerster, R. and Miguel, E. (2012). Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan. *The Quarterly Journal of Economics*, **127** (4), 1755–1812.

Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, **329** (5996), 1194–1197.

— (2011). An Experimental Study of Homophily in the Adoption of Health Behavior. *Science*, **334** (6060), 1269–1272.

Chen, K.-P. and Yang, S.-Z. (2002). Strategic Voting in Open Primaries. *Public Choice*, **112** (1-2), 1–30.

Chen, Y. (2009). Political Competition in Two-Stage Elections. *Working Paper*.

Clark, T. (1959). Internal Operation of the United States Supreme Court. *Journal of the American Judicature Society*, **43** (2), 45.

Coate, S. and Loury, G. (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *The American Economic Review*, **83** (5), 1220–1240.

Coleman, J. (1971). Internal Processes Governing Party Positions in Elections. *Public Choice*, **11** (1), 35–60.

Cooper, A. and Munger, M. C. (2000). The (Un)predictability of Primaries with Many Candidates: Simulation Evidence. *Public Choice*, **103** (3-4), 337–355.

Cornes, R. (1993). Dyke Maintenance and Other Stories: Some Neglected Types of Public Goods. *The Quarterly Journal of Economics*, **108** (1), 259–271.

— and Hartley, R. (2007a). Aggregative Public Good Games. *Journal of Public Economic Theory*, **9** (2), 201–219.

— and — (2007b). Weak links, Good Shots and Other Public Good Games: Building on BBV. *Journal of Public Economics*, **91** (9), 1684–1707.

Costa-Gomes, M. and Crawford, V. (2006). Cognition and Behavior in Two-Person Guessing Games: An Experimental Study. *The American Economic Review*, **96** (5), 1737–1768.

Crutzen, P. J. (2006). Albedo Enhancement by Stratospheric Sulfur Injections: A Contribution to Resolve a Policy Dilemma? *Climatic Change*, **77** (3), 211–220.

D'Arrigo, R., Wilson, R., Liepert, B. and Cherubini, P. (2008). On the Divergence Problem in Northern Forests: A Review of the Tree-Ring evidence and Possible Causes. *Global and Planetary Change*, **60** (3), 289–305.

De Nooy, W., Mrvar, A. and Batagelj, V. (2011). *Exploratory Social Network Analysis with Pajek*. Cambridge University Press.

Dell, M., Jones, B. F. and Olken, B. A. (2009). Temperature and Income: Reconciling New Cross-Sectional and Panel Estimates. *The American Economic Review Papers and Proceedings*, **99** (2), 198–204.

Deschênes, O. and Greenstone, M. (2011). Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US. *American Economic Journal: Applied Economics*, **3** (4), 152–185.

Downs, A. (1957). *An Economic Theory of Democracy*. New York: Harper and Row.

Duflo, E., Hanna, R. and Ryan, S. (2012). Incentives Work: Getting Teachers to Come to School. *The American Economic Review*, **102** (4), 1241–1278.

Edenhofer, O., Pichs-Madruga, R. and Sokona, Y. (eds.) (2011). *IPCC Expert Meeting on Geoengineering*, IPCC Working Group III Technical Support Unit, Potsdam Institute for Climate Impact Research.

Edmond, C. (2012). Information Manipulation, Coordination and Regime Change. *Working Paper*.

Epstein, L., Landes, W. and Posner, R. (2011). Why (and When) Judges Dissent: A Theoretical and Empirical Analysis. *Journal of Legal Analysis*, **3** (1), 101–137.

Fafchamps, M. and Quinn, S. (2012). Networks and Manufacturing Firms in Africa: Initial Results from a Randomised Experiment. *Working Paper*.

Feddersen, T. and Pesendorfer, W. (1996). The Swing Voter's Curse. *The American Economic Review*, **86** (3), 408–424.

Feely, R. A., Sabine, C. L., Lee, K., Berelson, W., Kleypas, J., Fabry, V. J. and Millero, F. J. (2004). Impact of Anthropogenic CO2 on the CaCO3 System in the Oceans. *Science*, **305** (5682), 362–366.

Fryer, R. (2007). Belief Flipping in a Dynamic Model of Statistical Discrimination. *Journal of Public Economics*, **91** (5), 1151–1166.

Gans, J. S. and Smart, M. (1996). Majority Voting with Single-Crossing Preferences. *Journal of Public Economics*, **59** (2), 219–237.

Genz, A. (2004). Numerical Computation of Rectangular Bivariate and Trivariate Normal and t Probabilities. *Statistics and Computing*, **14** (3), 251–260.

Gildea, J. (1990). Explaining FOMC Members' Votes. In T. Mayer (ed.), *The Political Economy of American Monetary Policy*, Cambridge University Press, pp. 211–227.

Goes, M., Tuana, N. and Keller, K. (2011). The Economics (or Lack Thereof) of Aerosol Geoengineering. *Climatic Change*, **109** (3-4), 719–744.

Goldin, C. and Rouse, C. (2000). Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians. *The American Economic Review*, **90** (4), 715–741.

Gole, T. (2011). Voters Who Care About Policy and Voters Who Care About Purity: A Citizen-Candidate Model of Primary Elections. *Working Paper*.

Goodell, J. (2010). *How to Cool the Planet*. Boston, Houghton Mifflin Harcourt.

Grieco, P. L. (2011). Discrete Games with Flexible Information Structures: An Application to Local Grocery Markets. *Working Paper*.

Haan, M. (2000). Endogenous Party Formation in a Model of Representative Democracy. In *Econometric Society World Congress 2000 Contributed Papers*, 0598, Econometric Society.

Hanna, R. N. and Linden, L. L. (2012). Discrimination in Grading. *American Economic Journal: Economic Policy*, **4** (4), 146–168.

Hansson, I. and Stuart, C. (1984). Voting Competitions with Interested Politicians: Platforms Do Not Converge to the Preferences of the Median Voter. *Public Choice*, **44** (3), 431–441.

Harrison, G. (2011). Randomisation and Its Discontents. *Journal of African Economies*, **20** (4), 626–652.

— and List, J. (2004). Field Experiments. *Journal of Economic Literature*, **42** (4), 1009–1055.

Havrilesky, T. and Schweitzer, R. (1990). A Theory of FOMC Dissent Voting with Evidence from the Time Series. In T. Mayer (ed.), *The Political Economy of American Monetary Policy*, Cambridge University Press, pp. 197–210.

Heal, G. (2009). The Economics of Climate Change: a Post-Stern Perspective. *Climatic Change*, **96** (3), 275–297.

— and Park, J. (2013). Feeling the Heat: Temperature and the Wealth of Nations, forthcoming.

Heckman, J. and Smith, J. (1995). Assessing the Case for Social Experiments. *The Journal of Economic Perspectives*, **9** (2), 85–110.

Hinich, M. J. and Ordeshook, P. C. (1969). Abstentions and Equilibrium in the Electoral Process. *Public Choice*, **7** (1), 81–106.

Hirshleifer, J. (1983). From Weakest-Link to Best-Shot: The Voluntary Provision of Public Goods. *Public Choice*, **41** (3), 371–386.

Holden, R. and Hummel, P. (2011). Optimal Primaries. *Working Paper*.

Hope, C. (2006). The Marginal Impact of CO2 from PAGE2002: An Integrated Assessment Model Incorporating the IPCC's Five Reasons for Concern. *Integrated Assessment*, **6** (1).

Hotelling, H. (1929). Stability in Competition. *Economic Journal*, **XXXIX**, 41–57.

Hsiang, S. M. and Narita, D. (2012). Adaptation to Cyclone Risk: Evidence from the Global Cross-section. *Climate Change Economics*, **3** (02).

Hummel, P. (2010). Flip-flopping from Primaries to General Elections. *Journal of Public Economics*, **94** (11), 1020–1027.

Iaryczower, M., Shi, X. and Shum, M. (2013). Words Get in the Way: The Effect of Deliberation in Collective Decision-Making. *Working Paper*.

— and Shum, M. (2012). The Value of Information in the Court: Get it Right, Keep it Tight. *The American Economic Review*, **102** (1), 202–237.

Intergovernmental Panel on Climate Change (2007). Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge University Press Cambridge.

IPCC (1996). Climate Change 1995: Impacts, Adaptation and Mitigation of Climate Change: Scientific-Technical Analysis. Contribution of Working Group II to the Second Assessment Report of the Intergovernmental Panel on Climate Change.

Itaya, J.-i., De Meza, D. and Myles, G. D. (1997). In Praise of Inequality: Public Good Provision and Income Distribution. *Economics Letters*, **57** (3), 289–296.

Jackson, M. O., Mathevet, L. and Mattes, K. (2007). Nomination Processes and Policy Outcomes. *Quarterly Journal of Political Science*, **2** (1).

Jung, A. (2011). An International Comparison of Voting by Committeees. *European Central Bank Working Paper No 1383*.

Keane, M. (2010). Structural vs. Atheoretic Approaches to Econometrics. *Journal of Econometrics*, **156** (1), 3–20.

Keith, D. W. (2000). Geoengineering the Climate: History and Prospect. *Annual Review of Energy and the Environment*, **25** (1), 245–284.

—, Parson, E. and Morgan, M. G. (2010). Research on Global Sun Block Needed Now. *Nature*, **463** (7280), 426–427.

Kemp, M. C. (1984). A Note of the Theory of International Transfers. *Economics Letters*, **14** (2), 259–262.

Keynes, J. (1936). *The General Theory of Interest, Employment and Money*. London: Macmillan.

Kimhi, A. (1994). Quasi Maximum Likelihood Estimation of Multivariate Probit Models: Farm Couples' Labor Participation. *American Journal of Agricultural Economics*, **76** (4), 828–835.

Lang, K. and Lehman, J.-Y. K. (2012). Racial Discrimination in the Labor Market: Theory and Empirics. *Journal of Economic Literature*, **50** (4), 959–1006.

Lavy, V. (2008). Do Gender Stereotypes Reduce Girls' or Boys' Human Capital Outcomes? Evidence from a Natural Experiment. *Journal of Public Economics*, **92** (10), 2083–2105.

Levy, G. (2005). Careerist Judges and the Appeals Process. *RAND Journal of Economics*, **36** (2), 275–297.

— (2007). Decision Making in Committees: Transparency, Reputation, and Voting Rules. *The American Economic Review*, **97** (1), 150–168.

Li, H., Rosen, S. and Suen, W. (2001). Conflicts and Common Interests in Committees. *The American Economic Review*, **91** (5), 1478–1497.

Liepert, B. G., Feichter, J., Lohmann, U. and Roeckner, E. (2004). Can Aerosols Spin Down the Water Cycle in a Warmer and Moister World? *Geophysical Research Letters*, **31** (6).

List, J. (2004). The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field. *The Quarterly Journal of Economics*, **119** (1), 49–89.

Maddison, D. and Rehdanz, K. (2011). The Impact of Climate on Life Satisfaction. *Ecological Economics*, **70** (12), 2437–2445.

Manne, A., Mendelsohn, R. and Richels, R. (1995). MERGE: A Model for Evaluating Regional and Global Effects of GHG Reduction Policies. *Energy Policy*, **23** (1), 17–34.

Matthews, H. D. and Caldeira, K. (2007). Transient Climate-Carbon Simulations of Planetary Geoengineering. *Proceedings of the National Academy of Sciences*, **104** (24), 9949–9954.

McClellan, J., Keith, D. W. and Apt, J. (2012). Cost Analysis of Stratospheric Albedo Modification Delivery Systems. *Environmental Research Letters*, **7** (3), 034019.

McGhee, E., Masket, S., Shor, B. and McCarty, N. (2013). A Primary Cause of Partisanship? Nomination Systems and Legislator Ideology. Available at SSRN: http://ssrn.com/abstract=1674091 or http://dx.doi.org/10.2139/ssrn.1674091.

Mendelsohn, R., Morrison, W., Schlesinger, M. E. and Andronova, N. G. (2000). Country-specific Market Impacts of Climate Change. *Climatic Change*, **45** (3-4), 553–569.

Mirhosseini, M. R. (2007). Primaries with Strategic Voters: Trading off Electability and Ideology. Ph.D. thesis chapter.

Mohan, J. E., Ziska, L. H., Schlesinger, W. H., Thomas, R. B., Sicher, R. C., George, K. and Clark, J. S. (2006). Biomass and Toxicity Responses of Poison Ivy (Toxicodendron radicans) to Elevated Atmospheric CO2. *Proceedings of the National Academy of Sciences*, **103** (24), 9086–9089.

Moore, J., Jevrejeva, S. and Grinsted, A. (2010). Efficacy of Geoengineering to Limit 21st Century Sea-level Rise. *Proceedings of the National Academy of Sciences*, **107** (36), 15699–15703.

Moreno-Cruz, J. B. (2010). Mitigation and the Geoengineering Threat. *Working Paper*.

Moro, A. and Norman, P. (2004). A General Equilibrium Model of Statistical Discrimination. *Journal of Economic Theory*, **114** (1), 1–30.

Morris, S. and Shin, H. (1998). Unique Equilibrium in a Model of Self-fulfilling Currency Attacks. *The American Economic Review*, **88** (3), 587–597.

— and — (2006). Heterogeneity and Uniqueness in Interaction Games. In L. E. Blume and S. N. Durlauf (eds.), *The Economy as an Evolving Complex System, III: Current Perspectives and Future Directions*, Oxford University Press, pp. 207–242.

— and Shin, H. S. (2003). Global Games: Theory and Applications. *Econometric Society Monographs*, **35**, 56–114.

Nordhaus, W. D. (2006). Geography and Macroeconomics: New Data and New Findings. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (10), 3510–3517.

— (2010). Economic Aspects of Global Warming in a post-Copenhagen Environment. *Proceedings of the National Academy of Sciences*, **107** (26), 11721–11726.

— and Boyer, J. (2000). *Warming the World: Economic Models of Global Warming*. MIT Press (MA).

— and Yang, Z. (1996). A Regional Dynamic General-Equilibrium Model of Alternative Climate-Change Strategies. *The American Economic Review*, **86** (4), 741–765.

Oak, M. P. (2006). On the Role of the Primary System in Candidate Selection. *Economics & Politics*, **18** (2), 169–190.

O'Brien, K., Sygna, L. and Haugen, J. E. (2004). Vulnerable or Resilient? A Multi-scale Assessment of Climate Impacts and Vulnerability in Norway. *Climatic Change*, **64** (1-2), 193–225.

Oman, L., Robock, A., Stenchikov, G. L. and Thordarson, T. (2006). High-latitude Eruptions Cast Shadow over the African Monsoon and the Flow of the Nile. *Geophysical Research Letters*, **33** (18).

Orcutt, G. H. and Orcutt, A. G. (1968). Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes. *The American Economic Review*, **58** (4), 754–772.

Osborne, M. J. and Slivinski, A. (1996). A Model of Political Competition with Citizen-Candidates. *The Quarterly Journal of Economics*, **111** (1).

Owen, G. and Grofman, B. (2006). Two-stage Electoral Competition in Two-Party Contests: Persistent Divergence of Party Positions. *Social Choice and Welfare*, **26** (3), Springer—569.

de Paula, A. and Tang, X. (2012). Inference of Signs of Interaction Effects in Simultaneous Games with Incomplete Information. *Econometrica*, **80** (1), 143–172.

Phelps, E. (1972). The Statistical Theory of Racism and Sexism. *The American Economic Review*, **62** (4), 659–661.

Poutvaara, P. (2003). Party platforms with Endogenous Party Membership. *Public Choice*, **117** (1-2), 79–98.

Price, J. and Wolfers, J. (2010). Racial Discrimination Among NBA Referees. *The Quarterly Journal of Economics*, **125** (4), 1859–1887.

Rabl, A. and van der Zwaan, B. (2009). Cost-Benefit Analysis of Climate Change Dynamics: Uncertainties and the Value of Information. *Climatic Change*, **96** (3), 313–333.

Rasch, P. J., Tilmes, S., Turco, R. P., Robock, A., Oman, L., Chen, C.-C. J., Stenchikov, G. L. and Garcia, R. R. (2008). An Overview of Geoengineering of Climate Using Stratospheric Sulphate Aerosols. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **366** (1882), 4007–4037.

Riboni, A. and Ruge-Murcia, F. (2011). Dissent in Monetary Policy Decisions. *Working Paper*.

Robock, A. (2000). Volcanic Eruptions and Climate. *Reviews of Geophysics*, **38** (2), 191–219.

— (2008). 20 Reasons Why Geoengineering May Be a Bad Idea. *Bulletin of the Atomic Scientists*.

—, Bunzl, M., Kravitz, B. and Stenchikov, G. L. (2010). A Test for Geoengineering? *Science*, **327** (5965), 530–531.

—, Marquardt, A., Kravitz, B. and Stenchikov, G. (2009). Benefits, Risks, and Costs of Stratospheric Geoengineering. *Geophysical Research Letters*, **36** (19).

—, Oman, L. and Stenchikov, G. L. (2008). Regional Climate Responses to Geoengineering with Tropical and Arctic SO2 injections. *Journal of Geophysical Research: Atmospheres (1984–2012)*, **113** (D16).

Roemer, J. E. (1997). Political Economic Equilibrium when Parties Represent Constituents: the Unidimensional Case. *Social Choice and Welfare*, **14** (4), 479—502.

— (2004). Indeterminacy of Citizen-Candidate Equilibrium. *Yale School of Management Working Papers*.

— (2006). *Political Competition: Theory and Applications*. Harvard University Press.

Rogoff, K. (1985). The Optimal Degree of Commitment to an Intermediate Monetary Target. *The Quarterly Journal of Economics*, **100** (4), 1169–1189.

— (1990). Equilibrium Political Budget Cycles. *The American Economic Review*, **80** (1), 21–36.

Ross, A. and Matthews, H. D. (2009). Climate Engineering and the Risk of Rapid Climate Change. *Environmental Research Letters*, **4** (4), 045103.

Samuelson, P. A. (1954). The Pure Theory of Public Expenditure. *The Review of Economics and Statistics*, **36** (4), 387–389.

Sandler, T. and Vicary, S. (2001). Weakest-Link Publc Goods: Giving In-Kind or Transferring Money in a Sequential Game. *Economics Letters*, **74** (1), 71–75.

Saporiti, A. and Tohmé, F. (2006). Single-Crossing, Strategic Voting and the Median Choice Rule. *Social Choice and Welfare*, **26** (2), 363–383.

Schlenker, W. and Roberts, M. J. (2008). Estimating the Impact of Climate Change on Crop Yields: The Importance of Nonlinear Temperature Effects. NBER Working Paper No. w13799.

Secretariat of the Convention on Biological Diversity (2012). *Geoengineering in Relation to the Convention on Biological Diversity: Technical and Regulatory Matters,*. Tech. Rep. Technical Series No.66, , Montreal.

Sen, A. and Williams, B. (eds.) (1982). *Utilitarianism and Beyond*. Cambridge University Press.

Serra, G. (2011). Why Primaries? The Party's Tradeoff Between Policy and Valence. *Journal of Theoretical Politics*, **23** (1), 21–51.

Shearer, B. (2004). Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment. *Review of Economic Studies*, **71** (2), 513–534.

Shepherd, J. (ed.) (2009). *Geoengineering the Climate: Science, Governance and Uncertainty*. Royal Society.

Smit, B. and Wandel, J. (2006). Adaptation, Adaptive Capacity and Vulnerability. *Global Environmental Change*, **16** (3), 282–292.

Smith, J. B., Schellnhuber, H.-J., Mirza, M. M. Q., Fankhauser, S., Leemans, R., Lin, E., Ogallo, L., Pittock, B., Richels, R., Rosenzweig, C. *et al.* (2001). Vulnerability to Climate Change and Reasons for Concern: A Synthesis. *Climate Change*, pp. 913–967.

Stern, N. N. H. (2007). *The Economics of Climate Change: the Stern Review*. Cambridge University Press.

Su, C. (2012). Estimating Discrete-Choice Games of Incomplete Information: A Simple Static Example. *Working Paper*.

Tilmes, S., Müller, R. and Salawitch, R. (2008). The Sensitivity of Polar Ozone Depletion to Proposed Geoengineering Schemes. *Science*, **320** (5880), 1201–1204.

Todd, P. and Wolpin, K. (2006). Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility. *The American Economic Review*, **96** (5), 1384–1417.

Tol, R. S. (2002). Estimates of the Damage Costs of Climate Change. Part 1: Benchmark Estimates. *Environmental and Resource Economics*, **21** (1), 47–73.

— (2008). Climate, Development and Malaria: An Application of FUND. *Climatic Change*, **88** (1), 21–34.

— (2009). The Economic Effects of Climate Change. *The Journal of Economic Perspectives*, **23** (2), 29–51.

— and Yohe, G. W. (2007). The Weakest Link Hypothesis for Adaptive Capacity: An Empirical Test. *Global Environmental Change*, **17** (2), 218–227.

Trenberth, K. E. and Dai, A. (2007). Effects of Mount Pinatubo Volcanic Eruption on the Hydrological Cycle as an Analog of Geoengineering. *Geophysical Research Letters*, **34** (15), L15702.

United Nations (2004). World Population to 2300. Department of Economic and Social Affairs, Population Division, United Nations, ST/ESA/SER.A/236. (United Nations, New York).

Vaughan, N. E. and Lenton, T. M. (2011). A Review of Climate Geoengineering Proposals. *Climatic Change*, **109** (3-4), 745–790.

Veres-Ferrer, E. J. and Pavía, J. M. (2012). On the Relationship Between the Reversed Hazard Rate and Elasticity. *Statistical Papers*, pp. 1–10.

Vicary, S. (1990). Transfers and the Weakest-Link: An Extension of Hirshleifer's Analysis. *Journal of Public Economics*, **43** (3), 375–394.

— (2011). Public Goods and the Commons: A Common Framework. *Journal of Public Economic Theory*, **13** (1), 47–69.

— and Sandler, T. (2002). Weakest-Link Public Goods: Giving In-Kind or Transferring Money. *European Economic Review*, **46** (8), 1501–1520.

Visser, B. and Swank, O. (2007). On Committees of Experts. *The Quarterly Journal of Economics*, **122** (1), 337–372.

Warr, P. G. (1983). The Private Provision of a Public Good is Independent of the Distribution of Income. *Economics Letters*, **13** (2), 207–211.

Weitzman, M. (2012). A Voting Architecture for the Governance of Free-Driver Externalities, with Application to Geoengineering. NBER Working Paper No. 18622.

Weitzman, M. L. (2009). On Modeling and Interpreting the Economics of Catastrophic Climate Change. *The Review of Economics and Statistics*, **91** (1), 1–19.

Wittman, D. (1973). Parties as Utility Maximizers. *The American Political Science Review*, **67** (2), 490–498.

— (1977). Candidates with Policy Preferences: A Dynamic Model. *Journal of Economic Theory*, **14** (1), 180–189.

— (1983). Candidate Motivation: A Synthesis of Alternative Theories. *The American Political Science Review*, **77** (1), 142–157.

Yohe, G. and Tol, R. S. (2002). Indicators for Social and Economic Coping Capacity—Moving Toward a Working Definition of Adaptive Capacity. *Global Environmental Change*, **12** (1), 25–40.

Appendix A: Appendix to Chapter 1

A.1 Proofs

A.1.1 Proof of Proposition 1

It is sufficient for judge i to have a unique cutoff x_i^* that the difference in utility between $a_i = 0$ and $a_i = 1$ is monotonically increasing in x_i, holding fixed the cutoff points of the other judges.

In our setting, that difference is:

$$x_i^* + \left\{ \Phi_2 \left[-\alpha_j(x_i^*), -\alpha_k(x_i^*), \omega_{jk} \right] - \Phi_2 \left[\alpha_j(x_i^*), \alpha_k(x_i^*), \omega_{jk} \right] \right\} \delta_i$$
$$+ \left\{ \Phi_2 \left[-\alpha_j(x_i^*), \alpha_k(x_i^*), \omega_{jk} \right] - \Phi_2 \left[\alpha_j(x_i^*), -\alpha_k(x_i^*), \omega_{jk} \right] \right\} (\delta_{ij} - \delta_{ik}),$$

where

$$\alpha_j(x_i^*) = \frac{x_j^* - \mu_j - \rho_{ij}(x_i^* - \mu_i)}{\sqrt{1 - \rho_{ij}^2}},$$
$$\alpha_k(x_i^*) = \frac{x_k^* - \mu_k - \rho_{ik}(x_i^* - \mu_i)}{\sqrt{1 - \rho_{ik}^2}}, \text{and}$$
$$\omega_{jk} = \frac{\rho_{jk} - \rho_{ij} \cdot \rho_{ik}}{\sqrt{\left(1 - \rho_{ij}^2\right) \cdot \left(1 - \rho_{ik}^2\right)}}.$$

The derivative of this difference with respect to x_i is:

$$1+ \delta_i \cdot \sqrt{1-\omega_{jk}^2} \cdot \left[\frac{\rho_{ij}}{\sqrt{1-\rho_{ij}^2}} \cdot \phi\left(\alpha_j(x_i^*)\right) + \frac{\rho_{ik}}{\sqrt{1-\rho_{ik}^2}} \cdot \phi\left(\alpha_k(x_i^*)\right) \right]$$
$$+ \left(\delta_{ij} - \delta_{ik}\right) \cdot \sqrt{1-\omega_{jk}^2} \cdot \left[\frac{\rho_{ij}}{\sqrt{1-\rho_{ij}^2}} \cdot \phi\left(\alpha_j(x_i^*)\right) - \frac{\rho_{ik}}{\sqrt{1-\rho_{ik}^2}} \cdot \phi\left(\alpha_k(x_i^*)\right) \right]$$

Without loss of generality, assume $\delta_{ij} \geq \delta_{ik}$. It is then sufficient that:

$$1 - \sqrt{1-\omega_{jk}^2} \cdot \frac{\rho_{ik}}{\sqrt{1-\rho_{ik}^2}} \cdot \phi\left(\alpha_k(x_i^*)\right) \cdot \left(\delta_{ij} - \delta_{ik} - \delta_i\right) > 0.$$

Substituting in for ω_{jk}, and assuming the largest possible value for $\alpha_k(x_i^*)$, gives:

$$\left(\delta_{ij} - \delta_{ik} - \delta_i\right) < \frac{\left(1-\rho_{ik}^2\right)}{\rho_{ik}} \cdot \sqrt{\frac{2\pi \cdot \left(1-\rho_{ij}^2\right)}{1-\rho_{ij}^2-\rho_{ik}^2-\rho_{jk}^2+2\rho_{jk}\cdot\rho_{ij}\cdot\rho_{ik}}}.$$

Noting that the RHS must be positive, it is sufficient for this condition that $\delta_i \geq \delta_{ij}, \delta_{ik} \geq 0$.

■

A.1.2 Proof of Proposition 2

Our proof of Proposition 2 is an extension of the proof in Morris and Shin (2006) to 3 players. The proof requires some additional notation not included in the main text.

Step 1: Existence of a threshold strategy This follows from Proposition 1.

We now restrict our attention to threshold strategies. Let $u_i(a_i, \Gamma_i(x_j^*, x_k^*, x_i), x)$ be judge i's expected payoff if their action is a_i, $\Gamma_i(x_j^*, x_k^*, x_i)$ is their belief about the other two judge's actions given they believe the other judges to be following cutoffs of x_j^*, x_k^* and they observe signal x_i, and their signal is x. A strategy profile is triple $\mathbf{x} = (x_i^*, x_j^*, x_k^*)$.

Define $\Pi_i(\Gamma_i(x_j^*, x_k^*, x_i), x_i) = u(1, \Gamma_i(x_j^*, x_k^*, x_i), x_i) - u(0, \Gamma_i(x_j^*, x_k^*, x_i), x_i)$. Then x is an equilibrium iff $x_i > x_i^* \iff \Pi_i(\hat{\Gamma}_i(s_j, s_k, x_i), x_i) > 0$ for all i and x_i.

Note that the problem set up in this way has the following properties (which resemble those in Morris and Shin (2006)):

- **Strategic Complementarities** If Γ dominates Γ', then $u_i(1, \Gamma, x_i) - u_i(0, \Gamma, x_i) \geq u_i(1, \Gamma', x_i) - u_i(0, \Gamma', x_i)$. This follows from $\delta_i \geq \delta_{ij}, \delta_{ik} \geq 0$.

- **Limit Dominance** There are signals sufficiently high and low that judges find it a dominant strategy give to the debate to the underdog and the favorite, regardless of Γ. These dominance regions are $x_i < -\delta_i$ and $x_i > \delta_i$.

- **Uniformly Positive** $(\underline{\kappa})$ **Sensitivity to the State** There exists $\underline{\kappa}$ such that if $x \geq x'$

$$[u(1,\Gamma,x) - u(0,\Gamma,x)] - [u(1,\Gamma,x') - u(0,\Gamma,x')] \geq \underline{\kappa}(x - x')$$

  This holds for $\underline{\kappa} \geq 1$.

- **Uniformly Bounded** $(\bar{\kappa})$ **Sensitivity to Opponents' Actions**: There exists $\bar{\kappa}$ such that if

$$[u(1,\Gamma,x) - u(0,\Gamma,x)] - [u(1,\Gamma',x) - u(0,\Gamma',x)] \geq \bar{\kappa}|\Gamma - \Gamma'|$$

  where

$$|\Gamma - \Gamma'| = sup|\Gamma(x_j^*, x_k^*, x_i) - \Gamma'(x_j^*, x_k^*, x_i)|$$

  This holds for $\bar{\kappa} \geq 2\delta_i$, noting that $\delta_i \geq \delta_{ij}, \delta_{ik} \geq 0$.

- **Stochastically Ordered Marginals** The conditional CDF of x_j given x_i, is increasing in x_i for all x_j. This follows from assuming $\rho_{ij} > 0 \; \forall \; i, j$.

- **Uniformly Bounded** (δ) **Marginals on Differences**: there exists $\nu > 0$ such that for all x and the 2x1 vector Δ,

$$\frac{d}{dx} F_i(\boldsymbol{x_i} + \Delta | x_i) \leq \nu$$

  This holds for $\nu \geq = \sqrt{\frac{1-\omega_{jk}^2}{2\pi}} \left[\sqrt{\frac{1-\rho_{ij}}{1+\rho_{ij}}} + \sqrt{\frac{1-\rho_{ik}}{1+\rho_{ik}}} \right]$.

Step 2: Existence of largest ($\bar{\mathbf{x}}$) and smallest ($\underline{\mathbf{x}}$) pure strategy profiles that satisfy iterated deletion of dominated strategies. This follows from on limit dominance, strategic complementarities, conditional state monotonicity and stochastically ordered marginals.

- By limit dominance, there exist regions where judges play some action with certainty.

- Just below that threshold, judges rule out opponents' strategies that involve not playing those actions should they receive signals above the dominance regions.

- For judges sufficiently close to the threshold, this then makes them have dominant strategies to vote the same way. This comes from strategic complementarities and stochastically ordered marginals.

- We can then eliminate strategies one by one, iterating this process.

More generally, this is a standard argument in supermodular games, as noted by Morris and Shin (2006).

Step 3: If $\underline{\kappa} > \bar{\kappa}\nu$ those largest and smallest strategy profiles are the same Suppose $\bar{\mathbf{x}} \neq \underline{\mathbf{x}}$. Then translate the cutoffs of $\underline{\mathbf{x}}$ left until each judge's cutoff lies to the left of their cutoff under $\bar{\mathbf{x}}$, but that one translated cutoff is equal under the two profiles. Let z be the amount of the translation, and without loss of generality let player i with type $\hat{\mathbf{x}}_i$ be the player whose cutoff is the same under the two profiles. Write $\tilde{\mathbf{x}}$ for the translated strategy profile.

We therefore have $\tilde{x}_j = \underline{x}_j + z$ for all j, and $\bar{x}_i = \tilde{x}_i$.

Then note:

$$\begin{aligned}\Pi_i(\hat{\Gamma}_i(\underline{x}_j, \underline{x}_k, \hat{x}_i + z), \hat{x}_i + z) &\geq \Pi_i(\hat{\Gamma}_i(\underline{x}_j, \underline{x}_k, \hat{x}_i + z), \hat{x}_i) + \underline{\kappa} z \\ &\geq \Pi_i(\hat{\Gamma}_i(\tilde{x}_j, \tilde{x}_k, \hat{x}_i), \hat{x}_i) + \underline{\kappa} z - \bar{\kappa}\nu z \\ &\geq \Pi_i(\hat{\Gamma}_i(\bar{x}_j, \bar{x}_k, \hat{x}_i), \hat{x}_i) + (\underline{\kappa} - \bar{\kappa}\nu)\, z\end{aligned}$$

If $\underline{\kappa} - \bar{\kappa}\nu > 0$, this implies that $\bar{\mathbf{x}}$ is not an optimal strategy. We thus a contradiction and $\bar{\mathbf{x}} = \underline{\mathbf{x}}$.

Therefore, a sufficient condition for a unique equilibrium is $\underline{\kappa} > \bar{\kappa}\nu$. Expanding and re-arranging gives Proposition 2:

$$\delta_i < \sqrt{\frac{\pi}{2(1-\omega_{jk}^2)}} \cdot \left(\sqrt{\frac{1-\rho_{ij}}{1+\rho_{ij}}} + \sqrt{\frac{1-\rho_{ik}}{1+\rho_{ik}}} \right)^{-1}$$

■

A.1.3 Proof of Proposition 3

Start by considering i's decision where $\delta_j = \delta_k = 0$, and therefore $x_j^* = x_k^* = 0$. If $\delta_i = 0$, then $x_i^* = 0$ straightforwardly.

At $x_i^* = 0$, with $x_j^* = 0$, note that if $\mu_j - \rho_{ij}\mu_i < 0$, $\alpha_j(x_i^*) = \dfrac{x_j^* - \mu_j - \rho_{ij}(x_i^* - \mu_i)}{\sqrt{1-\rho_{ij}^2}} < 0$, and similarly if $\mu_k - \rho_{ik}\mu_i$, $\alpha_k(x_i^*) < 0$. In that case, $\Phi_2\left[-\alpha_j(x_i^*), -\alpha_k(x_i^*), \omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i^*), \alpha_k(x_i^*), \omega_{jk}\right] > 0$. It follows that if δ_i becomes positive, x_i^* must become negative for the equation that defines x_i^* to hold. This holds for any value of δ_i.

Next, consider increasing δ_{ij} and δ_{ik} to be greater than zero as well. We then need to consider the sign of the second and third parts of the equation that defines x_i^*:

$$\left\{\Phi_2\left[-\alpha_j(x_i^*), -\alpha_k(x_i^*), \omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i^*), \alpha_k(x_i^*), \omega_{jk}\right]\right\}\delta_i$$
$$+\left\{\Phi_2\left[-\alpha_j(x_i^*), \alpha_k(x_i^*), -\omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i^*), -\alpha_k(x_i^*), -\omega_{jk}\right]\right\}(\delta_{ij} - \delta_{ik})$$

Note that while the fact that $\alpha_j(_i^*)$ and $\alpha_k(_i^*)$ are negative ensures the first difference is positive, it does not guarantee that the second difference is positive. *However*, it does guarantee that the first difference is larger in magnitude than the second difference, noting that the standard bivariate normal is symmetric.

Since we have assumed $\delta_i > \delta_{ij} > \delta_{ik} > 0$, it follows that the entire sum is positive. and so x_i^* is negative provided $\mu_n - \rho_{in}\mu_i > 0 \; \forall \; n \in j, k$.

Finally, consider the case where $\delta_j > 0$ and $\mu_n - \rho_{jn}\mu_j > 0 \; \forall \; n \in i, k$. By the argument above, $x_j^* < 0$. Because x_j^* enters positively in $\alpha_j(x_i^*)$, this makes $\alpha_j(x_i^*)$ more negative, and so the arguments above still hold: if $\mu_n - \rho_{in}\mu_i > 0 \; \forall \; n \in j, k$, $x_i^* \leq 0$. This result extends to the case where both δ_j and δ_k are greater than 0.

■

A.1.4 Proof of Proposition 4

We step through this proof for δ_i, the steps of the proof are analogous for δ_{ij} and δ_{ik}.

The fact that the signs of the total derivatives and the partial derivatives are the same is guaranteed by Proposition 2 and the uniqueness of equilibrium.[1]

Proposition 3 shows that at $\delta_i = 0$, the sign of the partial derivative is equal to the sign of $\Phi_2\left[-\alpha_j(x_i), -\alpha_k, \omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i), \alpha_k(x_i), \omega_{jk}\right]$ evaluated at the cutoff of $x_i^* = 0$.

Next, note that because δ_i only enters linearly into the equilibrium condition, the magnitude of the second derivative with respect to δ_i is proportional to the magnitude of the first derivative. While we cannot sign the second derivative, we can however conclude that it cannot change the sign of the first derivative for δ_i such that Proposition 2 holds.[2]

Finally, note that Proposition 3 states that if $\mu_m - \rho_{mn}\mu_n > 0 \; \forall \; m, n \in i, j, k$, then $\Phi_2\left[-\alpha_j(x_i), -\alpha_k, \omega_{jk}\right] - \Phi_2\left[\alpha_j(x_i), \alpha_k(x_i), \omega_{jk}\right]|_{x_i=0} > 0$ and $\delta_{ij} = \delta_{ik}$ guarantees that there

[1]The steps of this proof are very similar to the proof of Proposition 2, and have been omitted. They are available from the authors at request.

[2]For $\delta_i > 0$, note that strategic complementarities guarantee that the other players' cutoffs move in the same direction as player i's. However, the sign of the second derivative depends not on the cross-partials, but on the sign of $\frac{\partial x_j^*}{\partial x_i} - \rho_{ij}$. Since ρ_{ij} is only constrained to lie between 0 and 1, this implies that the second derivative can be positive or negative.

are no strategic interactions between individual pairs of players that can overturn the direct effect of this difference at any level of δ_i. The difference is therefore negative for all values of δ_i (subject to Proposition 2), and x_i^* is monotonically decreasing in δ_i.

■

A.1.5 Proof of Proposition 5

The model is identified if we can solve for the vector $(\rho_{12}, \rho_{13}, \rho_{23}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3, \delta_1, \delta_2, \delta_3)$ if we observe the following eight conditional probabilities:

$$\begin{aligned}
&\Pr(a_{1c} = 0, a_{2c} = 0, a_{3c} = 0 \,|\, R_c, D_{1c}, D_{2c}, D_{3c});\\
&\Pr(a_{1c} = 1, a_{2c} = 0, a_{3c} = 0 \,|\, R_c, D_{1c}, D_{2c}, D_{3c});\\
&\qquad\qquad\vdots\\
&\Pr(a_{1c} = 1, a_{2c} = 1, a_{3c} = 1 \,|\, R_c, D_{1c}, D_{2c}, D_{3c}).
\end{aligned}$$

Step 1: identification of $(\rho_{12}, \rho_{13}, \rho_{23}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$. Consider the eight conditional probabilities for the case in which $D_{1c} = D_{2c} = D_{3c} = 0$. In that case, it is trivial that $x_{1c}^* = x_{2c}^* = x_{3c}^* = 0$. Then the conditional probabilities are wholly determined by a trivariate probit, and identification of $(\rho_{12}, \rho_{13}, \rho_{23}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$ is straightforward and well understood: see Ashford and Sowden (1970).

By way of illustration, note that β_i and γ_i are identified by the probability of $a_i = 1$, conditional on R_c:

$$\Pr(a_i = 1 \,|\, R_c, D_{1c} = D_{2c} = D_{3c} = 0) = \Phi\left(\beta_i \cdot R_c + \gamma_i \cdot R_c^2\right). \tag{A.1}$$

The parameters $(\rho_{12}, \rho_{13}, \rho_{23})$ are identified through correlations in observed outcomes, conditional on R_c. (Note that these parameters can even be identified simply through pairwise correlations: see Kimhi (1994).)

Step 2: identification of $(\delta_1, \delta_2, \delta_3)$. Consider the eight conditional probabilities for the three cases in which $D_i = 1; D_j = 0 \;\forall\; j \neq i$. Note that, in those three cases, $\delta_{ic} = \delta_i$ and $\delta_{jc} = 0$, and therefore, $x_{jc}^* = 0$. Then Proposition 4 shows that there is a one-to-one relationship between δ_i and x_{ic}^*. Then we are done if we can identify x_i^*. We can identify x_i^* from the probability that $a_i = 1$, treating β_i and γ_i as known and choosing any fixed value for R_c:

$$\Pr(a_i = 1 \,|\, R_c, D_{ic} = 1, D_{jc} = 0) = \Phi\left(\beta_i \cdot R_c + \gamma_i \cdot R_c^2 - x_i^*\right). \tag{A.2}$$

■

A.2 Regression results on heterogeneous effects

Table A.1 allows the effect of lagged dissent to vary by judge gender. We find no significant difference between men and women.

Table A.2 tests whether dissent with individual judges matters (as distinct from dissenting with *both* peers). (Note that the variable '`Dissented`' in equation 1.1 refers to dissenting from both peers, because a judge who dissents from only one peer remains in the majority.) In column (1), we show that the effect of past dissent can be wholly explained by the effect of having dissented from a senior colleague; however, the difference between the effect of dissenting with a senior colleague and a junior colleague is not significant.

In Table A.3, we interact lagged dissent with the debate round, in order to test whether effects vary over the course of the the tournament. We find significant heterogeneity in the effect of dissenting on the probability of voting for the favorite: we estimate that a judge who dissents in Round 1 is about 27 percentage points more likely to vote for the favorite in Round 2 (*i.e.* $0.368 - 0.051 \times 2$), whereas the effect has essentially dissipated by Round 7. We find no analogous heterogeneity in the effect of past dissent on whether a judge subsequently dissents.[3] These results are intriguing: they suggest that the cost of dissent may decline as a tournament continues (perhaps because tournament organizers learn more about judge quality), or that judges learn more about their own ability (so that dissents hold less informational value to judges).

[3]Note, though, that we *do* estimate a significant heterogeneous effect by tournament round of having been dissented *against*.

Table A.1: *Regression results: Heterogeneity by gender*

	(1)	(2)	(3)	(4)
	Votes for the favorite		**Dissents**	
	Male	**Female**	**Male**	**Female**
Dummy: Just dissented	0.147	0.034	-0.185	-0.122
	(0.054)***	(0.066)	(0.034)***	(0.040)***
Dummy: Just dissented against	0.012	0.093	-0.029	0.002
	(0.045)	(0.040)**	(0.027)	(0.029)
Distance (senior)	-0.004	-0.003	0.002	-0.004
	(0.004)	(0.004)	(0.003)	(0.003)*
Distance (junior)	-0.003	-0.004	0.001	0.000
	(0.003)	(0.003)	(0.002)	(0.002)
Dummy: Did not judge	0.027	-0.055	-0.003	-0.086
	(0.072)	(0.075)	(0.059)	(0.055)
Judge × tournament dummies	✓	✓	✓	✓
Round × tournament dummies	✓	✓	✓	✓
Committees	603	603	603	603
Observations	990	819	990	819
R^2	0.011	0.011	0.025	0.020
Gender equality tests (p-values):				
Dummy: Just dissented		0.184		0.241
Dummy: Just dissented against		0.161		0.455
Distance (senior)		0.775		0.105
Distance (junior)		0.879		0.787

Standard errors in parentheses. Errors are clustered by judge and by committee.
Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table A.2: *Regression results: Heterogeneity by dissenting peer*

	(1) Votes for the favorite	(2) Dissents
Dummy: Dissented with both peers	**-0.010**	**-0.139*****
	(0.082)	**(0.054)**
Dummy: Dissented with the senior peer	**0.107***	**-0.018**
	(0.055)	**(0.036)**
Dummy: Dissented with the junior peer	**0.005**	**-0.006**
	(0.054)	**(0.028)**
Dummy: Did not judge	0.104***	-0.039
	(0.038)	(0.027)
Distance (senior)	-0.008*	-0.000
	(0.004)	(0.002)
Distance (junior)	-0.002	0.001
	(0.003)	(0.002)
Judge $\times$ tournament fixed effects	✓	✓
Committees	603	603
Observations	1809	1809
R^2	0.179	0.189
H_0: Equality, dissent with senior and junior peers (p)	0.179	0.800

Standard errors in parentheses. Errors are clustered by judge and by committee.
Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

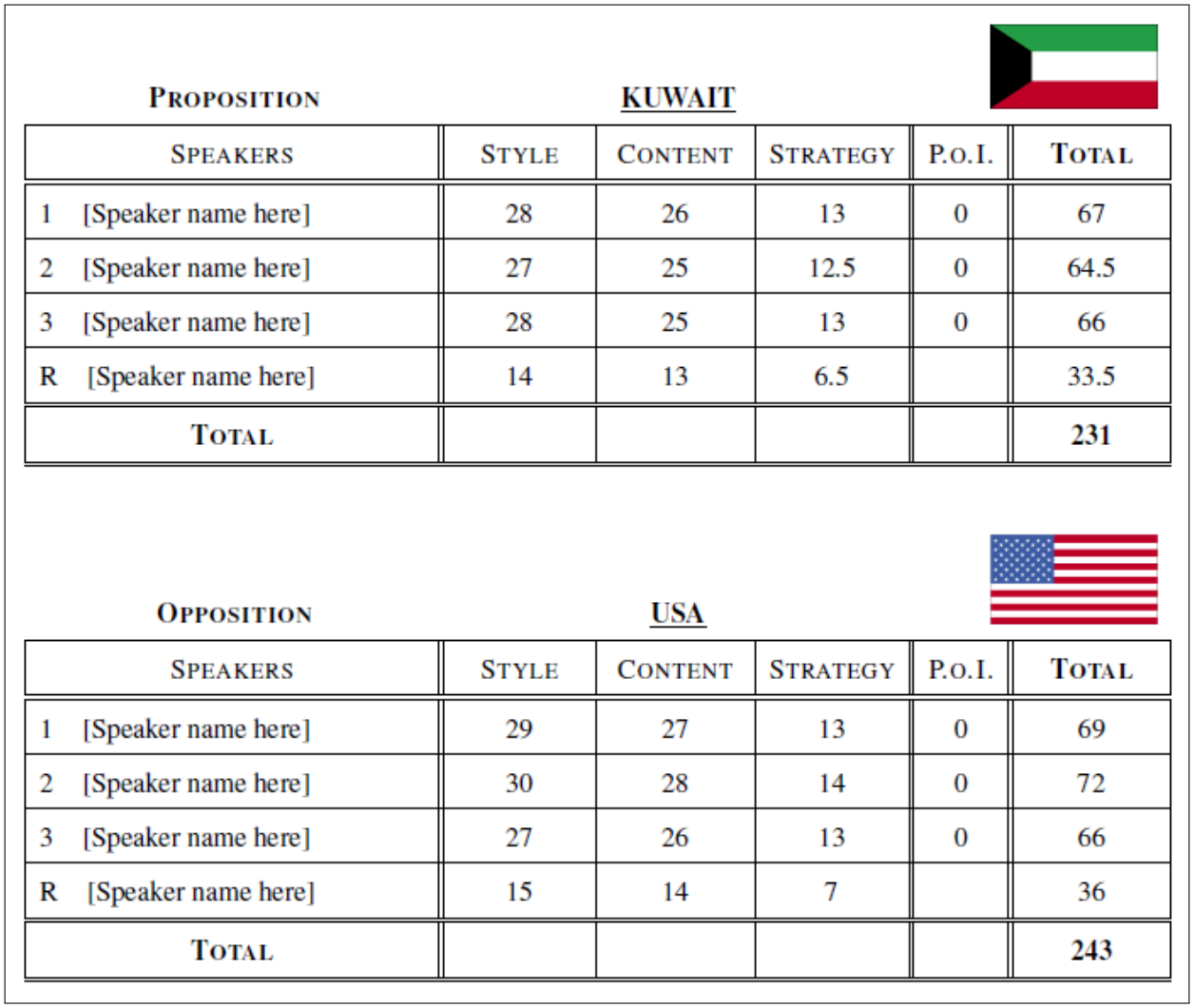
Table A.3: *Regression results: Heterogeneity by tournament round*

	(1) **Votes for the favorite**	(2) **Dissents**
Dummy: Just dissented	0.354 (0.102)***	-0.240 (0.063)***
Just dissented × Round	-0.051 (0.020)**	0.015 (0.013)
Dummy: Just dissented against	0.127 (0.093)	-0.149 (0.051)***
Just dissented against × Round	-0.016 (0.017)	0.026 (0.010)***
Distance (senior)	-0.004 (0.003)	-0.001 (0.002)
Distance (junior)	-0.003 (0.002)	0.001 (0.002)
Dummy: Did not judge	-0.012 (0.051)	-0.036 (0.041)
Judge × tournament dummies	✓	✓
Round × tournament dummies	✓	✓
Committees	603	603
Observations	1809	1809
R^2	0.012	0.025

Standard errors in parentheses. Errors are clustered by judge and by committee.
Confidence: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

A.3 An example ballot

Figure A.1 shows an example of a judge's ballot, taken from a debate between Kuwait and the United States. This judge voted for the United States (by a margin of 243 marks to 231). Note that the total score is simply the sum of the individual speaker totals, and that each individual speaker total is simply the sum of marks for the categories 'Style', 'Content', 'Strategy', and 'Points of Information' ('P.o.I'). 'Style' refers to the way that a speaker presents: for example, whether the the speech is delivered at an appropriate speed, and whether the speaker makes eye contact with the audience. 'Content' refers to the substance of the arguments: for example, whether arguments are logical, and substantiated with persuasive evidence. 'Strategy' refers to the speaker's identification of the key issues in the debate, and the consistency of the speaker's material with the material of his or her teammates. 'Points of Information' are short interjections that speakers are permitted to make during their opponents' speeches.



PROPOSITION — **KUWAIT**

SPEAKERS	STYLE	CONTENT	STRATEGY	P.O.I.	TOTAL
1 [Speaker name here]	28	26	13	0	67
2 [Speaker name here]	27	25	12.5	0	64.5
3 [Speaker name here]	28	25	13	0	66
R [Speaker name here]	14	13	6.5		33.5
TOTAL					**231**

OPPOSITION — **USA**

SPEAKERS	STYLE	CONTENT	STRATEGY	P.O.I.	TOTAL
1 [Speaker name here]	29	27	13	0	69
2 [Speaker name here]	30	28	14	0	72
3 [Speaker name here]	27	26	13	0	66
R [Speaker name here]	15	14	7		36
TOTAL					**243**

Figure A.1: *An example ballot from one judge*

A.4 Dynamic Structural Model

In this appendix, we sketch a structural model which takes into account the potentially dynamic nature of dissent aversion in a finite-horizon game. We assume a 'super-utility' function that aggregates over (i) the utility a judge gains from expressing his or her preferences across multiple rounds, and (ii) the disutility from dissenting across multiple rounds. Without loss of generality, we can solve this using a value function, in which we treat the disutility as accruing at the end of the final round.

Denote each round by r, with the total number of rounds being $T = 8$. Let V_r be the *expected utility from all future play, entering round r*. This will be a function of δ_i, which is i's choice variable in entering the stage game. Let V_T be the *terminal value function*, being the value that is subtracted after period R as a cost of having dissented. We suppress the state variable(s) for simplicity.

Behavior in the stage game is as discussed in the body of this paper. We now nest that stage game in this dynamic game by allowing judge i to *choose* her value for δ_i before she plays each round.

Start by considering period $T + 1$, after the final round T. Let κ_d be the *cost* of reaching the end of the game having dissented d times (we normalize the payoff from having never dissented at zero). We stack these values in the vector $\boldsymbol{\kappa}$. Let s_i be i's stock of dissents, where this is a scalar. We can then write V_{T+1}:

$$V_{T+1}(s_i) = \begin{cases} -\kappa_1 & \text{if } s_i = 1; \\ -\kappa_2 & \text{if } s_i = 2; \\ -\kappa_3 & \text{if } s_i = 3; \\ -\kappa_4 & \text{if } s_i = 4; \\ -\kappa_5 & \text{if } s_i = 5; \\ -\kappa_6 & \text{if } s_i = 6; \\ -\kappa_7 & \text{if } s_i = 7; \\ -\kappa_8 & \text{if } s_i = 8. \end{cases} \tag{A.3}$$

Next, consider the final round, T. After judge i has observed the ranking difference R (which generates the μ parameters, the mapping of which is estimated using the results in the committees of the final round where no judge has dissented, as in Proposition 5), s_j, s_k, and i's own signal $x_{i,T}$, the expected utility in round T from voting for the favorite is:

$$EU_T(a_i = 1|s_i, T, s_j, s_k, x_{i,T}) = x_{i,T} + P(a_j = a_k = 1|x_{i,T})V_T(s_i) + P(a_j = a_k = 0|x_{i,T})V_T(s_i + 1)$$

The expected utility from voting for the underdog is:

$$EU_T(a_i = 0|s_i, T, s_j, s_k, x_{i,T}) = P(a_j = a_k = 1|x_{i,T})V_T(s_i + 1) + P(a_j = a_k = 0|x_{i,T})V_T(s_i)$$

As usual, indifference between the two pins down $x^*_{i,T}$:

$$0 = x^*_{i,T} + \left[P(a_j = a_k = 1|x^*_{i,T}) - P(a_j = a_k = 0|x^*_{i,T})\right](V_T(s_i) - V_T(s_i + 1))$$

Note that if $V_{T+1}(s_i)$ is linear, then this model is isomorphic to the model in the body of the paper with $\delta_{ij} = \delta_{ik} = 0$.

Since we have a unique $x^*_{i,T}$ conditional on R, s_j, s_k and $x_{i,T}$, we then simply integrate out over each of those. Begin with $x_{i,T}$ (note that once we integrate over $x_{i,T}$ the probabilities

of the other two players' actions become unconditional):

$$
\begin{aligned}
&\mathbb{E}\left(U_T(s_i, R, s_j, s_k)\right) \\
&= \Pr(a_i = 1) \cdot \mathbb{E}(x_{iT} \,|\, a_i = 1) \\
&\quad + \left[(1 - P(a_i = 1, a_j = a_k = 0) - P(a_i = 0, a_j = a_k = 1))\right] V_T(s_i) \\
&\quad + \left[P(a_i = 1, a_j = a_k = 0) + P(a_i = 0, a_j = a_k = 1)\right] V_T(s_i + 1) \\
&= \Pr(a_i = 1) \cdot \mathbb{E}(x_{iT} \,|\, a_i = 1) + V_T(s_i) \\
&\quad + \left[\Pr(a_i = 1, a_j = a_k = 0) + \Pr(a_i = 0, a_j = a_k = 1)\right] \cdot (V_T(s_i + 1) - V_T(s_i)) \\
&= \Pr(x_{iT} > x^*_{iT}) \cdot \mathbb{E}(x_{iT} \,|\, x_{iT} > x^*_{iT}) + V_T(s_i) + \Big[\Pr(x_{iT} > x^*_{iT}, x_{jT} < x^*_{jT}, x_{kT} < x^*_{kT}) \\
&\quad + \Pr(x_{i,R} < x^*_{i,R}, x_{j,R} > x^*_{j,R}, x_{k,R} > x^*_{k,R})\Big] \cdot (V_T(s_i + 1) - V_T(s_i)) \\
&= \phi(x^*_{i,R} - \mu_{i,R}) + \left[1 - \Phi(x^*_{i,R} - \mu_{i,R})\right] \cdot \mu_{i,R} + V_T(s_i) \\
&\quad + \Big[\Phi_3(x^*_{i,R} - \mu_{i,R}, -(x^*_{j,R} - \mu_{j,R}), -(x^*_{k,R} - \mu_{k,R}), \mathbf{\Omega}) \\
&\quad + \Phi_3(-(x^*_{i,R} - \mu_{i,R}), x^*_{j,R} - \mu_{j,R}, x^*_{k,R} - \mu_{k,R}, \mathbf{\Omega})\Big] \cdot (V_T(s_i + 1) - V_T(s_i))
\end{aligned}
$$

Since this is a function of the remaining conditional parameters (R, s_j, s_k), we can then integrate out over them, using the empirical distributions R_T, S_k and S_j, rather than equilibrium predictions. Taking this approach guarantees equilibrium uniqueness.

$$
\begin{aligned}
EU_T(s_i) = &\sum^{r \in R_T} P(R = r) \sum^{s_j \in S_{j,T}} P(s_j = s_j) \sum^{s_k \in S_{k,T}} P(s_k = s_k) \\
&\Big\{\phi(x^*_{i,T} - \mu_{i,T}) + (1 - \Phi(x^*_{i,T} - \mu_{i,T}))\mu_{i,T} + V_T(s_i) \\
&- \Big[\Phi_3(x^*_{i,T} - \mu_{i,T}, -(x^*_{j,T} - \mu_{j,T}), -(x^*_{k,T} - \mu_{k,T}), \Omega) \\
&+ \Phi_3(-(x^*_{i,T} - \mu_{i,T}), x^*_{j,T} - \mu_{j,T}, x^*_{k,T} - \mu_{k,T}, \Omega)\Big] (V_T(s_i) - V_T(s_i + 1))\Big\} \\
= &\, V_T(s_i)
\end{aligned}
$$

We have suppressed the conditionality of the μs and the cutoffs for simplicity. This

expression can be evaluated using numerical integration over three parameter spaces, for the value of arriving in Round T with s_i dissents.

A similar approach allows us to create $V_{T-1}(s_i)$, and so on back through the rounds. This will eventually collapse back to:

$$0 = x^*_{i,0} + \left[P(a_j = a_k = 1|x^*_{i,0}) - P(a_j = a_k = 0|x^*_{i,0})\right](V_1(0) - V_1(1))$$

We therefore have expressions which generate the equilibrium cutoffs for each player in each round, as a function of the distribution parameters $(\rho_{12}, \rho_{13}, \rho_{23}, \beta_1, \beta_2, \beta_3, \gamma_1, \gamma_2, \gamma_3)$ and the $\boldsymbol{\kappa}$ vector.

We can then estimate these parameters in a similar fashion to the model in the body of the paper. We estimate the mapping of R to the μs, and the ρs, by imposing $\kappa_1 = 0$ (analogous to assuming $\delta_0 = 0$ in the body of the paper). Then in round 8 there will be debates in which $s_i = s_j = s_k = 0$, and so all players do not care about coordination. This locks $x_i = x_j = x_k = 0$, and so we can estimate the distributional parameters, before moving on to estimating κ.

The twist here is estimation of the $\boldsymbol{\kappa}$ vector, having estimated the distribution parameters. We proceed in the following fashion. Pick a candidate $\boldsymbol{\kappa}$ vector. Solve for cutoffs for Round 8, and then solve for the Value Function at Round 8 by integrating over the empirical parameter distributions. Use $V_8(s_i)$ to solve for cutoffs in Round 7, and so on back to Round 1. We then iterate to converge on an estimated $\boldsymbol{\kappa}$ vector.

Note that with sufficient data we could estimate the $\boldsymbol{\kappa}$ vector from just Round 8. However, since it is unlikely that we would observe all possible sets of $\{s_i, s_j, s_k\}$, we can use earlier rounds to get at higher values of κ. There will be debates in which κ_n is still a possible outcome, and therefore still decision relevant, but $\kappa_{>n}$ is not, so we can uniquely pin down

κ_n. Then move back a round to get at κ_{n+1} and so on.

In this fashion, it is possible to analyze, identify and estimate a dynamic version of the model used in the body of this paper. This estimation is more computationally difficult, and so is currently work in progress.

Appendix B: Appendix to Chapter 2

B.1 A Continuous Model of Public Gobs

The model outlined in this appendix is closely based on the work of Cornes and Hartley (2007a).

B.1.1 Model

There are n countries. Country i's preferences are represented by the utility function $u_i = u_i(x_i, G)$, where x_i is the quantity of national private consumption and G is the total quantity of the public gob. We impose the following assumptions:

- *Well behaved preferences:* for all i, the function $u_i(.)$ is everywhere strictly quasiconcave in both arguments and everywhere continuous. It is strictly increasing in x_i, but it is *not*, however, strictly increasing in G.
- *Marginal utility of G must become non-positive:* for all i, there exists some $\tilde{G}_i$ such that $\frac{\partial u_i(.)}{\partial G} \leq 0 \; \forall \; G > \tilde{G}_i$.
- *Linear individual budget constraints:* Country i's budget constraint requires that $x_i + c_i q_i \leq m_i$ where $q_i \geq 0$ is her contribution to the public gob. The unit cost c_i and income m_i are strictly positive and exogenous.

- *Summative aggregation function:*[1] the total supply of the public gob is the sum of all individual contributions $G = \sum_{j=1}^{n} q_j = q_i + G_{-i}$ where G_{-i} is the contributions made by all countries except i.

- *Normality:* For every country i, both the private good and public good are weakly normal.

Our model is identical to Cornes and Hartley (2007a) in all regards except two: we have not restricted the marginal utility of the public good to be positive (the first assumption) and we have assumed that the marginal utility eventually becomes negative for all countries (the second assumption).[2] As we will see, this does not affect the existence of a unique equilibrium, but does generate an interesting comparison between the non-cooperative outcome and the socially optimal outcome.

The budget constraint defined above can be rewritten to explicitly include the contributions of others: $x_i + c_i G \leq m_i + c_i G_{-i}$. As Cornes and Hartley (2007a) note, this requires that i can only consume a bundle that does not exceed her full income ($M_i = m_i + C_i G_{-i}$), and where her private consumption does not exceed her private income ($x_i \leq m_i$).

Countries choose non-negative values of x_i and q_i to maximize their utility subject to their budget constraint and a given level of G_{-i}. Given well behaved utility functions, for any non-negative level of G_{-i} there exists a unique utility-maximizing contribution level $\hat{q}_i$ (which may be zero). By varying G_{-i}, we generate best response functions $\hat{q}_i = b_i(G_{-i})$. A Nash equilibrium consists of every country selecting the level of contribution that is the best response to the choices of all of the other countries.

---

[1] For more on aggregation functions, see Appendix B.2.

[2] The awkward double meaning of the word "good" is confusing in this context. We conform to the traditional language except where necessary to avoid confusion.

B.1.2 Noncooperative Outcome

Following Cornes and Hartley (2007a), we now define each country's replacement function $r_i(G, m_i, c_i)$. This is the amount that country i contributes in equilibrium if the aggregate level of the gob is G, their income is m_i and their cost is c_i. Cornes and Hartley (2007a) show that this function has the form:

$$r_i(G, m_i, c) \equiv \max\left\{\frac{m_i - \xi_i^{-1}(G)}{c_i} + G, 0\right\}$$

Where $\xi_i^{-1}(G)$ is the inverse of $\xi_i(M_i)$, country i's demand for the public good as a function of their full income M_i. Cornes and Hartley (2007a) show that this function is defined for all $G \geq \underline{G}_i$, where $\underline{G}_i$ is country i's "standalone value", the amount that they would contribute if no other country was contributing at all ($r_i(\underline{G}_i, m_i, c) = \underline{G}_i$). They also show that it is well behaved, in the sense that it is continuous, everywhere non-increasing in G and strictly decreasing in G wherever it is positive.

These individual replacement functions can be used to define an aggregate replacement function $R(G)$:

$$R(G, \mathbf{m}, \mathbf{c}) = \sum_{j=1}^{n} r_j(G, m_j, c_j)$$

where $\mathbf{m} \equiv (m_1, m_2, ..., m_n)$ and $\mathbf{c} \equiv (c_1, c_2, ..., c_n)$. Cornes and Hartley (2007a) show that this aggregate replacement function inherits the properties of the individual replacement functions: it is continuous, everywhere non-increasing and strictly decreasing where positive, and it is defined for all $G \geq \max\{\underline{G}_1, \underline{G}_2, ..., \underline{G}_n\}$.

The aggregate replacement function $R(G)$ leads to a simple characterization of Nash equilibrium: a Nash equilibrium is a strategy profile $\hat{\mathbf{q}}$ such that $\hat{q}_j = r_j(G^n, m_j, c_j)$, for

$j = 1, 2, ...n$ where $G^n = \sum_{j=1}^{n} \hat{q}_j$. Cornes and Hartley (2007a) show that in their setting, there exists a unique Nash equilibrium.

Our modification of the model of Cornes and Hartley (2007a) does not substantively affect their proofs of the existence of any of the individual replacement functions, the aggregate replacement function and a unique Nash equilibrium. The only amendment required is to change, under certain circumstances, the sets for which an individual country is up against a corner solution of not contributing at all. In Cornes and Hartley (2007a), nations are up against a corner solution because the marginal utility of consumption at the point where the nation privately consumes all of their income, and does not contribute to the public gob, exceeds the marginal utility of the gob. In our setting it is possible that the marginal utility of the gob may be negative for an individual country, and so they will be against a corner solution regardless of their level of private consumption. This does not substantively affect the analysis.

We therefore conclude that the continuous public gob game has a unique Nash equilibrium, which is determined by the distribution of income, cost functions and preferences over private consumption and the public gob. We can further state that if there is a common cost c such that $c_i = c \ \forall i$, that the level of public gob in equilibrium is falling in that cost. At the limit, as $c \to 0$, $G \to \max\{\tilde{G}_1, \tilde{G}_2, ...\tilde{G}_n\}$, and the level of public gob goes to the amount at which the nation most desirous of the gob receives the maximum amount of utility from it. In this sense, our model nests the model of Weitzman (2012).

B.1.3 Comparison with Socially Optimal Outcome

We can compare the non-cooperative equilibrium of the public gob game with the socially optimal level. The social optimum is characterized as it is in the standard public goods game:

$$G^* = G \text{ s.t. } \sum_{j=1}^{n} \frac{\partial u_j(y_j, G)}{\partial G} = \frac{\partial u_i(x_i, G)}{\partial x_i} c \; \forall \; i$$

Essentially, the social planner equates the marginal utility of an extra unit of the gob, being the sum of all of the individual marginal utilities, with the marginal cost of an extra unit, which is lowering the consumption of a specific country. The social planner therefore equates the marginal utilities of each of the individual countries. Country specific cost parameters complicate this analysis slightly, since we then have to ask *which* countries do the production and consider the question of transfers, but do not change it dramatically.[3]

Given general utility functions, we cannot give a clean comparison of the non-cooperative and social planner's outcomes. We can, however, make the following points. Firstly, note that for $c = 0$, the non-cooperative outcome sets the marginal utility of the individual most desirous of the gob to 0, while the social planner sets the sum of marginal utilities of the public gob to 0. Since the marginal utilities of all other nations are negative at the level of the non-cooperative outcome, this implies that at $c = 0$ the social planner selects a level of the gob below that of the non-cooperative outcome. By continuity of the utility functions and the aggregate replacement function, this is also true for very low levels of c.

Next, consider very high levels of c. By assuming Inada-like conditions to guarantee that the marginal utilities of the nations from the gob start positive, and do not decrease too quickly, then there will be a point in which the non-cooperative outcome is such that all nations still have a positive marginal utility of the gob in equilibrium, even if it is lower than their marginal utility from additional consumption and so they do not contribute any further. At this point, we have a traditional free-rider dynamic, and the non-cooperative outcome equates private costs and benefits, while the social planner equates the sum of benefits with the private cost of production, and therefore selects a higher level of the public

[3]Note that Cornes and Hartley (2007a) do not consider the social planner's problem.

gob. Therefore, at sufficiently high levels of c, the socially optimal outcome exceeds the non-cooperative outcome.

Since both sets of outcomes are continuous in c, the intermediate value theorem implies there will be a point at which the two outcomes are equal.[4] Whether a public gob features a free-driver dynamic, as discussed in this paper, or a more traditional free-rider dynamic, is therefore contingent on the cost of providing the good. We might then consider public goods games a subset of public gob games, in which the cost of production is sufficiently high that the possibility of negative marginal utility from the public good is simply never encountered.

B.1.4 Implications for Heterogeneity and Redistribution

What are the implications of this continuous model for heterogeneity and redistribution? Firstly, note that the wedge between the non-cooperative and social planner's outcome is related to heterogeneity of preferences, income and costs. In the extreme case, if all nations have the same preferences, then the social planner's outcome always exceeds the non-cooperative outcome, as there is no case in which one nation has a positive marginal utility from the public gob without all nations also preferring greater levels of it. If $c = 0$, then the non-cooperative outcome is socially optimal, as all nations select the same level of the good. Free-driving requires preference heterogeneity, while free-riding does not.

Secondly, an important factor for the existence of a free-driver dynamic is the relationship between preferences, and income and costs. If the countries that are most desirous of the public gob also happen to be the countries with the highest incomes, or the lowest costs of production, then it they will have higher replacement functions, and so a free-driver dynamic becomes more likely. On the other hand, if the nations most desirous of the gob

[4]With general functional forms, we cannot immediately guarantee that there will be a unique crossing, however. We leave characterization of the requirements of single-crossing for future research.

are also the poorest, or least likely to be able to effectively deploy the technology to produce it, then the tradeoff between the gob and private consumption will lead those countries to opt for lower levels of the gob, potentially heading off any chance of free-driving. Given the discussion in Section 2.3, it seems likely that in the context of geoengineering, we are in the second category: that the countries who are most likely to pursue geoengineering are also the countries that can least afford it. This leans against the possibility of free-driving, although the relative affordability of the technology, as discussed in the paper, suggests that we cannot rely on financial cost as a sufficient means of avoiding a dangerous clash of national and global priorities.

The standard result with respect to redistribution in public goods games is Bergstrom *et al.* (1986), who show that if the set of contributors is held fixed, income redistribution does not affect the level of the good supplied. This result also holds in the public gob game described here, for the same reasons as in Bergstrom *et al.* (1986) and in Cornes and Hartley (2007a): agents equate marginal utility from consumption and the public gob, and so redistribution of income simply leads nations to re-equate them at the same level of the public gob. This result relies on common costs across nations: if costs vary across nations, then redistribution from a high cost to a low cost nation (within the set of contributors) increases the level of the public good.[5]

If redistribution occurs from non-contributors to contributors, then in equilibrium this will increase the level of the public good provided. This unambiguously increases the utility of the contributors (who, by virtue of being contributors, have positive marginal utility from the gob), but the effect on the utility of the non-contributors is unclear, as Cornes and Hartley (2007a) show. It depends on the tradeoff between the lost private utility and the change in utility from the boost in the public gob. If the public product is a pure good, then when there are many non-contributors and relatively few contributors, it is possible

[5]For more, see Cornes and Hartley (2007a).

for non-contributors to gain. Redistribution may help ameliorate the free-rider problem. In a free-driver setting, the possibility that the non-contributors in fact receive negative utility from the increase in the public gob means that redistribution *exacerbates* the free-driver problem. While we must be very careful about making claims in this area, this points to a difficulty with transfers as a means of avoiding the free-driver problem. If nations are unable to commit, then transfers to the nations most desirous of geoengineering may have the effect of loosening their budget constraints, and making them more likely to geoengineer, not less.

Consider, on the the other hand, the possibility of in-kind transfers, such as assistance with adaptation. Such transfers, rather than loosening the budget constraint, would change nations' preferences over geoengineering, by lowering the impacts of climate change. This would lead nations pursuing geoengineering to contribute less, and so ameliorate the free-driver dynamic. We may view such transfers as "preference redistribution" of a sort: in enables non-contributing countries to affect the preferences of contributing countries. Note again the connection to heterogeneity: this approach would be effective because it lowers variation in climate change damages, and therefore also lowers heterogeneity in desire to pursue geoengineering.

B.2 Aggregation Technologies for Public Goods, Inequality and Preferences

This appendix contains extends the continuous model of Appendix B.1 to different aggregation functions, and sketches a new taxonomy for public good, bad and gob games.

B.2.1 Aggregation Technologies

Standard models of public goods provision focus on an aggregation technology where the amount of the public good provided is the sum of each agent's individual contribution. Hirshleifer (1983) first drew attention to models of public goods production do not fall into this class. In particular, Hirshleifer (1983) contrasted three archetypal cases of public goods provision:

1. Summative: $G = \sum_{i \in I} g_i$
2. Best-shot: $G = max\ g_i$
3. Weakest link: $G = min\ g_i$

Hirshleifer (1983) gives the example of dyke building as a weakest-link good, and this set up has inspired a range of examples. Barrett (2007) uses this framework to discuss the provision of a range of global public goods, from asteroid defense as a best-shot situation to disease eradication as a weakest-link public good.

B.2.2 Free-Riding and Free-Driving Under Different Aggregation Technologies

The traditional analysis of free-riding considers a pure public good and summative aggregation. It is fairly straight forward to note that a best-shot aggregation function also features

free-riding: one actor provides the public good, and all other actors contribute nothing. In this case, we can identify the level of contribution associated with a given Nash equilibrium, as it is simply the stand alone value of the contributor. The analysis is slightly complicated by the fact that there will be a number of Nash equilibria (as Hirshleifer (1983) points out), one for each actor who is willing to contribute a sufficient amount that the actor most desirous for the public good does not decide to volunteer to contribute a greater amount, but this does not significantly complicate matters.

On the other hand, weakest-link aggregation functions imply very different behavior. In this case, the outcome is dependent on the behavior of the actor least desirous of the public good. Every other actor would prefer a larger level of the public good than the least desirous actor selects, but cannot volunteer to provide that level due to the aggregation technology. We therefore have the mirror image of the free-driving problem in the public gob game - in this case we have one agent selecting to underprovide the good, on behalf of the collective.

Our results concerning heterogeneity and inequality also follow this pattern. In the case of summative public goods, Warr (1983), followed by Kemp (1984) and Bergstrom *et al.* (1986), demonstrate an invariance result on the effect of redistribution and inequality. As long as the set of contributors to a public good does not change, redistribution of income between actors does not change the equilibrium level of provision good provision, with each agent changing their provision of the public good by the change in their income. An important implication of this result is that if all agents have the same demand function for the public good, then (subject to corner solutions) changes in inequality have no effect on the level of public good provision.[6]

Cornes (1993) demonstrates that this result is specific to a summative aggregation tech-

[6]If agents are at a corner solution, they are non-contributors to the public good. Equalizing income redistributions from contributors to non-contributors decreases the level of the public good.

nology. When the production technology is weakest-link, increases in wealth inequality *decreases* the provision of the public good. This occurs because the level of public goods provision is determined the actor who is least desirous of the public good or who can least afford it, and when inequality increases this actor becomes poorer still, and chooses a lower level of contribution. The difference between the socially optimal level of public good provision and the competitive levels therefore increases.[7]

Cornes (1993) does not go on to explicitly consider best-shot public goods. However, an analogous argument applies: for best-shot public goods, increases in inequality *increase* the level of public goods provision. For all of the pure strategy Nash equilibria, an increase in inequality increases both the level of public good provision, and in some cases increases in inequality may rule out the equilibria with the lowest levels of provision. We conclude then that in best-shot cases, increases in inequality lead to higher public goods provision and, subject to differences in the cost of provision, a smaller wedge between the optimal social provision and the competitive equilibrium.[8]

B.2.3 A Taxonomy

In Appendix B.1, we noted that the public gob game can take on both free-riding and free-driving behavior, depending on the parameters. It can also take on either nature depending on the aggregation technology: note that under best-shot aggregation the public gob game will naturally feature free-driving for a sufficiently low cost, and free-riding for a sufficiently high cost, while under weakest-link aggregation all nations will have positive marginal utilities for the public good in equilibrium, and so it will resemble a free-driving game for any (positive) level of cost.

[7] There is a literature that focuses on the possibility of Pareto-improving transfers. See, for example, Vicary (1990), Sandler and Vicary (2001) and Vicary and Sandler (2002).

[8] An analogous argument, made in the context of summative public goods, is Itaya *et al.* (1997).

We can go one step further, and extend this analysis to the modeling of "public bads", such as pollution. Typically, we approach the modeling of public bads, such as greenhouse gas emissions, that are costly to reduce, by reversing the production process and viewing them instead as the provision of a costly public good, such as clean air. We assume the existence of some arbitrarily large maximum amount of possible emissions, and view individual contributions to the public good as selecting some level of emissions below this maximum level.[9]

In the summative case, this reframing of the problem has little effect on the analysis. Subject to changes in functional forms (to preserve the appropriate curvature of functions), equilibrium levels of public goods provision, private consumption and utility are preserved. In the weakest-link and best-shot cases, however, the direct mapping does not occur. Instead, as might be expected, the production technologies are flipped: a best-shot public bad, in which the level of pollution is selected by the dirtiest actor, becomes a weakest-link public good. Similarly, a weakest-link public bad, in which the level of pollution is selected by the cleanest actor, becomes a best-shot public good. The effects of changes in inequality are also flipped.

Combining each of the elements of this discussion, we can generate the taxonomy in Table B.1. In this context, we can give a more general definition of free-riding and free-driving. A free-rider situation is one in which every actor would prefer a higher level of public good provision, but is not willing to privately provide an additional unit. All of the agents free-ride each others' contributions. On the other hand, there are situations in which every actor would prefer a higher level of public good provision and many of them would be willing to provide an additional unit, but are unable to due to the actions of another actor. In

[9] An alternative is to frame such problems as "commons" public goods situations, as opposed to "subscription" cases; see Vicary (2011).

this sense, the recalcitrant actor, who is holding the collective provision down, is a free-driver.

Table B.1: ***A Taxonomy of Aggregation Functions and Preferences***

	Public Good	Public Bad	Public Gob
Best Shot	Free Riding	Free Driving	Free Riding for high c, Free Driving for low c
Summative	Free Riding	Free Riding	Free Riding for high c, Free Driving for low c
Weakest Link	Free Driving	Free Riding	Free Driving

This taxonomy has the merit of directing our attention to the fact that the dynamics of a public good game depend on the combination of the aggregation function, the shape of preferences and the cost of public provision. It also highlights the relationship of the wedge between the non-cooperative outcome and the social planner's solution and the level of inequality, in either resources or preferences, within the relevant population of actors.

We suspect that these classifications can be extended to "better-shot" and "weaker-link" aggregation functions, as defined by Cornes and Hartley (2007b), by defining them either directly with reference to the difference between the marginal incentives of the relevant actor and the social planner, or indirectly by considering the effect of changes in inequality on the wedge between socially optimal provision and the level of competitive provision. We leave this for future research.

Appendix C: Appendix to Chapter 3

C.1 Proof of Lemma 1

This proof is due to Chen (2009).

Single-peakedness of $U(x_i, x_k)$ is equivalent to strict quasi-concavity. Since the summation in $U(x_i, x_k)$ preserves concavity, we can focus on an individual candidate from the right, and therefore the concavity of $ln[P(x_L, x_R)]$ and $ln\left[u\left(x_L, x_k\right) - u\left(r, x_k\right)\right]$.

$ln[P(x_L, x_R)]$ is concave by Assumption 2. Concavity of $ln\left[u\left(x_L, x_k\right) - u\left(x_R, x_k\right)\right]$ is guaranteed by concavity of the utility function and Assumption 1 as follows:

$$\frac{\partial^2 ln\left[u\left(x_L, x_k\right) - u\left(x_R, x_k\right)\right]}{\partial x^2} = \frac{u''(x_L, x_k)\left[u\left(x_L, x_k\right) - u\left(r, x_k\right)\right] - \left(u''(x_L, x_k)\right)^2}{\left[u\left(x_L, x_k\right) - u\left(x_R, x_k\right)\right]^2} < 0$$

Thus, $U(x_i, x_k)$ is single-peaked for $x_i \in [\underline{x}, L_r]$.

C.2 Proof of Lemma 2

Single-crossing requires, for all $k > j$, $\frac{\partial}{\partial a}U(k, x_R, a) - U(j, x_R, a) > 0$.

$$\begin{aligned}\frac{\partial}{\partial a}U(k, x_R, a) - U(j, x_R, a) &= \frac{\partial}{\partial a}\Big\{P(k, x_R)u(k, a) + (1 - P(k, x_R)u(r, a) \\ &\qquad - \left[P(j, x_R)u(j, a) + (1 - P(j, x_R)u(r, a)\right]\Big\} \\ &= P(k, x_R)\left[\frac{\partial u(k, a)}{\partial a} - \frac{\partial u(r, a)}{\partial a}\right] - P(j, x_R)\left[\frac{\partial u(j, a)}{\partial a} - \frac{\partial u(r, a)}{\partial a}\right]\end{aligned}$$

Note that $P(k, x_R) > P(j, x_R)$ by monotonicity of the $P()$ function, and $\left[\frac{\partial u(k,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}\right] > \left[\frac{\partial u(j,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}\right]$ by concavity of the utility function. However, $\left[\frac{\partial u(k,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}\right] < 0$, and so we cannot claim without further that the whole statement is positive.[1]

Instead, set the statement to be positive, and re-arrange to obtain:

$$\frac{P(k, x_R)}{P(j, x_R)} < \frac{\frac{\partial u(k,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}}{\frac{\partial u(j,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}}$$

$$\frac{P(k, x_R) - P(j, x_R)}{P(j, x_R)} < \frac{\frac{\partial u(k,a)}{\partial a} - \frac{\partial u(j,a)}{\partial a}}{\frac{\partial u(j,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}}$$

Consider this inequality as $j \to k$. Then:

$$\frac{\frac{\partial P(j, x_R)}{\partial j}}{P(j, x_R)} < \frac{-\frac{\partial^2 u(j,a)}{\partial a \partial j}}{\frac{\partial u(j,a)}{\partial a} - \frac{\partial u(r,a)}{\partial a}}$$

We therefore have that if the condition for monotonicity in Section 3.2 holds, then the single-crossing property holds in Section 3.3.

C.3 Proof of Proposition 1

(a) In order that no other citizen with the same preferred policy position wishes to enter, we need that the expected ego returns from entry are lower than the costs of entry, given that entry will not change the expected outcome of the general election. The expected ego returns are the product of the probability of winning the primary election $(\frac{1}{2})$, the probability of winning the general election $(P(x_L^*(\mu_L), r))$ and the ego returns from election victory (b). Further, if $P(x_L^*(\mu_L), r)b \leq 2c$, then there is an equilibrium in

[1] Note that this implies the related proof in Chen (2009) is incorrect.

which a citizen at the party median enters, since any entrant with a different position loses, and the withdrawal of the single candidate yields $-\infty$.

(b) If $P(x_L^*(\mu_L), r)b > c$, then it is worthwhile for a citizen located at μ_L to enter, because the expected ego returns are greater than the cost of entry, and since their position in the general is flexible they can only change the expected election outcomes in their favor. A candidate located at μ_L beats any other candidate in the primary election with certainty, so no other one-candidate equilibrium exists.

(c) If there is a single candidate from a position other than the party median, then any citizen whose expected general election outcome the median prefers can win the primary outcome by entering. Using the definition of such a successful challenger as $\tilde{x}$ in Part (i), Part (ii) states that such a challenger must prefer to remain out of the contest. This condition is also sufficient, because no candidate outside of $(\tilde{x}, x_c)$ can successful challenge, if any candidate inside $(\tilde{x}, x_c)$ prefers entry then so does $\tilde{x}$ (as their preferred policy position is further away from x_c's), and x_c does not prefer to exit as they would obtain $-\infty$ if they did so.

C.4 Proof of Proposition 2

(a) If $P(x_L^*(\mu_L), r)b > 2c$, then it is worthwhile for two candidates arbitrarily close to μ_L to remain in the primary to obtain the ego returns, even if their presence does not significantly affect the general election result. Entry at the location of an existing candidate results in a certain loss in the primary election, and so a two-player equilibrium exists with certainty for $P(x_L^*(\mu_L), r)b > 2c$.

(b) If the median voter is not indifferent between the candidates in a two-candidate equilibrium, then one of them loses the primary with certainty, and by Lemma 3 does not enter.

(c) The first part of (i) provides conditions so that no entrant can win: x st $U(x^*(s), r, x) =$

$U(x^*(x_m), r, x)$ is the voter who is indifferent between x_m and the challenger, while x st $U(x^*(s), r, x) = U(x^*(x_e), r, x)$ is the voter who is indifferent between x_e and the challenger. (i) states that the mass between these two points cannot be larger than the mass either to their left or right.

Alternatively, (ii) states that the mass between these two points is exactly equal to the mass to their left and right. In that case, we require that a potential entrant prefers the expected outcome from not entering

$$\frac{1}{2}P(x^*(x_m), y)[(u(x^*(x_m), s) - u(y, s)] - \frac{1}{2}P(x^*(x_e), y)[(u(x^*(x_e), s) - u(y, s)]$$

to the expected outcome from entry

$$\frac{1}{3}P(x^*(s), y)[u(x^*(s), s) - u(y, s) + b] - \frac{1}{3}P(x^*(x_m), y)[(u(x^*(x_m), s) - u(y, s)] - \frac{1}{3}P(x^*(x_e), y)[(u(x^*(x_e), s) - u(y, s)] - c$$

Finally, (iii) simply states that the two candidates cannot locate at the same point, else they are vulnerable to an entrant on either side.

C.5 Proof of Proposition 3

By Lemma 4 there is no equilibrium in which all three candidates have the same ideal point. By Lemma 3, the two cases described in Proposition 3 remain.

a *The positions of the candidates are not all the same, and each candidate receives one-third of the votes:* Part (i) simply states the conditions by which t_1 and t_2 are indifferent between the requisite candidates. This has to be the case for all three candidates to be potential winners. Part (ii) guarantees that x_1 does not prefer to exit. The utility from entry is:

$$\frac{1}{3}P(x^*(x_1), y)[b + u(x^*(x_1), x_1) - u(y, x_1)] + \frac{1}{3}P(x^*(x_2), y)[(u(x^*(x_2), x_1) - u(y, x_1)] + \frac{1}{3}P(x^*(x_3), y)[(u(x^*(x_3), x_1) - u(y, x_1)] - c$$

while the utility from exit is:

$$\frac{1}{2}P(x^*(x_2),y)[(u(x^*(x_2),x_1)-u(y,x_1)]+\frac{1}{2}P(x^*(x_3),y)[(u(x^*(x_3),x_1)-u(y,x_1)]$$

Note that by single-peaked preferences, we can make this necessary condition tighter by stating that $P(x^*(x_2),y)[(u(x^*(x_2),x_1)-u(y,x_1)] \geq P(x^*(x_3),y)[(u(x^*(x_3),x_1)-u(y,x_1)]$.

b *The positions of the candidates are all different, and the middle candidate obtains a smaller fraction of the votes than the other two, who receive equal vote share:* Part (i) again states the conditions by which t_1 and t_2 are indifferent between the requisite candidates. Part (ii) states that for candidate 2 the difference in utility between the other two candidates is greater than the cost of standing.

C.6 Proof of Proposition 4

In this proof, we follow the analogous proof in Osborne and Slivinski (1996), concluding with the case that does not match (hence $P(x^*(x_1),r)b > kc$ is not sufficient).

Let the number of candidates be n. By Lemma 3, candidates 1 and n are winners in the primary. First, suppose that $x_1 = x_2$. Then $x_3 > x_1$ by Lemma 4. 1 and 2 are winners by Lemma 3, and if 1 withdraws, 2 is the sole winner of the primary. The payoff to withdrawal is $P(x^*(x_1),r)[u(x^*(x_1),x_1)-u(r,x_1)]$, while the payoff to entry is strictly less than $P(x^*(x_1),r)[u(x^*(x_1),x_1)-u(r,x_1)]+\frac{1}{k}P(x^*(x_1),r)b-c$. For entry to be preferred, we therefore require that $P(x^*(x_1),r)b > kc$. A similar argument can be made that $a_{n-1} = a_n$ requires $P(x^*(x_n),r)b > kc$. Since $P(x^*(x_n),r) > P(x^*(x_1),r)$, the requirement is $P(x^*(x_1),r)b > kc$.

Now suppose that $x_1 < x_2$ and $x_{n-1} < x_n$. Since $x_2 \leq x_{n-1}$, either $\frac{x_1+x_n}{2} \geq x_2$ or $\frac{x_1+x_n}{2} \leq x_{n-1}$.

In the first case, candidate 1 prefers to withdraw unless $P(x^*(x_1),r)b > kc$.. If they do so, then all citizens previously voting for candidate 1 now vote for candidates at x_2, who

become the clear winners. Then the returns to entry are at most:

$$\frac{1}{k}P(x^*(x_1),y)[b+u(x^*(x_1),x_1)-u(y,x_1)]+\frac{1}{k}P(x^*(x_n),y)[(u(x^*(x_n),x_1)-u(y,x_1)] \\ +\frac{k-2}{k}P(x^*(x_2),y)[(u(x^*(x_2),x_1)-u(y,x_1)]-c$$

While the returns to exit are $P(x^*(x_2),y)[(u(x^*(x_2),x_1)-u(y,x_1)]$. Since the expected utility function is concave to the right of x, and $x^*(x')>x\ \forall\ x'>x$:

$$\frac{1}{2}P(x^*(x_1),y)[u(x^*(x_1),x_1)-u(y,x_1)]+\frac{1}{2}P(x^*(x_n),y)[(u(x^*(x_n),x_1)-u(y,x_1)] \\ <P(x^*(x_2),y)[(u(x^*(x_2),x_1)-u(y,x_1)]$$

and therefore withdrawal is preferable unless $P(x^*(x_1),r)b>kc$.

In the second case, $\frac{x_1+x_n}{2}\leq x_{n-1}$ implies $\frac{u(x^*(x_n),x_n)+u(x^*(x_1),x_n)}{2}<u(x^*(x_n),x_n)$ and $\frac{P(x^*(x_n),r)+P(x^*(x_1),r)}{2}<P(x^*(x_{n-1}),r)$. It follows that:

$$P(x^*(x_n),r)u(x^*(x_n),x_n)+P(x^*(x_1),r)u(x^*(x_1),x_n)-2P(x^*(x_2),r)u(x^*(x_2),x_n)<0$$

The difference between the returns to entry and exit are then:

$$\frac{1}{k}P(x^*(x_1),y)[b+u(x^*(x_1),x_1)-u(y,x_1)]+\frac{1}{k}P(x^*(x_n),y)[(u(x^*(x_n),x_1)-u(y,x_1)] \\ -\frac{2}{k}P(x^*(x_{n-1}),y)[(u(x^*(x_{n-1}),x_1)-u(y,x_1)]-c$$

However, since utility is *not* necessarily concave to the left of x, we cannot say that:

$$\frac{1}{2}P(x^*(x_1),y)[u(x^*(x_1),x_n)-u(y,x_n)]+\frac{1}{2}P(x^*(x_n),y)[(u(x^*(x_n),x_n)-u(y,x_n)] \\ <P(x^*(x_{n-1}),y)[(u(x^*(x_{n-1}),x_n)-u(y,x_n)]$$

And so this case does *not* necessarily require $P(x^*(x_n),r)b>kc$. If it did, then since $P(x^*(x_n),r)>P(x^*(x_1),r)$, a necessary condition would be $P(x^*(x_n),r)b>kc$, but we cannot make that statement.